Association Analysis of Additive Effects and Epistasis Between Human Candidate Malaria Protective Genes

Thesis

Association Analysis of Additive Effects and Epistasis between Human Candidate Malaria Protective Genes

Carolyne Mukami Ndila BSc, MSc.
The Open University

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
August 2015

Affiliated Research Centre
KEMRI-Wellcome Trust Research Centre, Kilifi, Kenya

Collaborating establishment
MalariaGEN Consortium, Oxford, United Kingdom

DATE of SUBMISSION : 22 JUNE 2015
DATE of AWARD : 11 AUGUST 2015

Declaration of Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: 

Date: 11/08/2015

Abstract

Malaria is a major cause of childhood death in Africa and host genetic factors play a key role in determining survival from this disease. Although many candidate loci have been identified, there have been difficulties in confirming the significance of some of these loci. To some extent this might be explained by the added complexity of epistasis, or gene-gene interactions. Through this thesis I aimed: (1) to re-appraise a range of candidate malaria-association genes using a large-scale case-control study of severe malaria (SM) in Kilifi, Kenya; (2) to compare different approaches to detecting epistatic interactions; (3) to look for evidence of epistasis between candidate genes in my data set; (4) to examine the haplotype structure and linkage disequilibrium (LD) patterns for two such implicated variants (HbS and α^+thalassaemia) and their gene regions, that co-exist in the Kilifi population, and (5) to use these exemplars as a starting point for investigating the process of detecting epistasis in SM in a genome-wide association study (GWAS). Out of 71 candidate genes investigated, I observed that polymorphisms affecting various aspects of red blood cells (including *HBB*, *HBA*, *G6PD*, *FREM3*, *INPP4B*, *ATP2B4* and *ABO*) were among those associated with the strongest signals of differential susceptibility to SM. Because of their prominence in malaria, HbS and α^+thalassaemia were used to illustrate interaction analysis at the GWAS level. This included looking at the structure of the genomic regions surrounding the genes. As expected, a single haplotype of approximately 200kb was seen surrounding HbS, which then diverged into 2 major haplotypes spanning a further 1Mb either side, an observation that was largely explained by ethnicity. In contrast, no marked LD/haplotype structure was observed in the genomic region surrounding the α^+thalassaemia deletion, suggesting that this is a very old polymorphism. Through this study, I confirmed the negative epistasis seen between HbS and α^+thalassaemia using a study design (case-control) that was different to that used previously (cohort), although this was not among the most significant of the interactions I detected. I searched for pairwise interactions between these two polymorphisms at a genome wide level using heterozygous and additive models for HbS and α^+thalassaemia respectively. For each scan a single region reaching a significance level of $<10^{-7}$ was found (*STX18* for HbS and *MYEOV* for α^+thalassaemia), plus several other novel signals were identified in the 10^{-6} to 10^{-7} significance region. Further work will be required to validate these signals and the challenge will be to try and understand their biological relevance. This is now becoming possible with datasets in many diseases, including malaria, being released into the public domain. But, as this Kenyan study has shown, having large group sizes, high quality clinical and genetic data, it is possible to begin to explore genetic interactions in a disease setting.

Acknowledgments

This work would not have been possible without the help and support of many people. To all of them, I am much obliged.

I am deeply grateful to my supervisors, Professor Thomas N. Williams (TW) and Dr Kirk A. Rockett, who have always been there for me throughout my Ph.D, with their wisdom and unwavering support without which this work would have been impossible to accomplish. TW has been an amazing role model for me, both as a scientist and as a mentor, who basically introduced me to malaria epidemiology. His deep insights both into science and human nature, implicit trust and willingness to make time for me and respond thoughtfully, no matter his schedule, have enabled me to achieve and tap on potential that I would not otherwise have realised. Kirk has been an indispensable source of support and guidance throughout the MalariaGEN Consortium projects and this thesis. He is generous with his time and ideas and has given me more than he ever needed to; moreover, our regular meetings have kept me going throughout this Ph.D.

I would also like to extend my deepest gratitude to Dr. Taane Clark for giving me a taste of real statistical genetics. Taane has mentored and encouraged me since I was the MalariaGEN data fellow, for which I am very grateful.

My work heavily relies on very good experimental data. Therefore, I would like to thank my collaborators Professor Dominic Kwiatkowski (DK) and the MalariaGEN Consortium, at the Wellcome Trust Centre for Human Genetics (WTCHG) for creating a large-scale data infrastructure. In particular, I am thankful to the lab team, Anna Richardson, Christina Hurbbart, Kate Rowlands, and all the other DK group members at the WTCHG and Wellcome Trust Sanger Institute (Team 112). My sincere gratitude also goes to the

administrative team, Victoria Cornelius (Vikki), Kimberly Johnson, Christa Heinrichs for their support and for making my stay in Oxford possible. I would like to mention Vikki in particular for your overwhelming support for me throughout this journey. I am thankful to the Statistics/analysis team, Chris Spencer, Gavin Band, Geraldine Clarke, Si Quang Le, George Busby, and Ellen Leffler for building my confidence in the statistics minefield. I am particularly indebted to Gavin for taking a direct interest in my work. My heartfelt goes to Jenny Shelton, you nonetheless found the time to support and encourage me.

To the IT/informatics team, many thanks for the wonderful systems that your collective effort put at my disposal. In particular, thank you Robert Hutton for your support and making sure that I was able to connect to the WTCHG Computer Clusters away from Oxford.

Along this path, I made some friends who are worth mentioning Vysaul, Manju, Lucas, Deus, Tobias, Edith your friendship and scientific advice brought humour and sanity as I did my work.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I was funded by the KEMRI-Wellcome Trust Strategic Award. In particular, I am most grateful to Dr Samson Kinyanjui, Liz Murabu, Keith Kipoto, Professor James Nokes, and Dr Francis Ndungu for coordinating my studies and making sure everything has gone smoothly all these four years. Special thanks go to Dr Pete Bull, Dr Abdirahman Abdi and Dr Evasius Bauni for being my Ph.D. committee members and offering guidance over the years. I am grateful to Professor James Kahindi for your support and mentorship.

Much of the work would not have been possible without the support of my close colleagues under TW team especially Alex, Johnstone, Metrine, Herbert, Oscar, Sophie, Mary, Prophet, Emily, George, Gideon, Mabibo, Jacob and Ruth.

I am also grateful to the good people of the Kilifi health demographic and surveillance system who availed themselves to be used in the case-control studies and all the other people whose samples were used.

Also, I would like to thank, Professor Julian Knight and Professor Lars Hviid for taking the time out of their busy schedules to be the referees of this thesis. I would also like to thank them for providing invaluable guidance in identifying areas of this thesis that required some revision.

Finally, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits, as well as my brother Dr. Videlis N. Nduba and my sister Dr. Janet Ndila for their support. And most of all to my loving, supportive, encouraging, and patient husband Dr. Said S. Ngala whose faithful support during the final stages of this Ph.D. is so much appreciated. Last but not the least, to my wonderful children who have cheered me up with their sweet smiles whenever I felt like I was too exhausted.

Attributions

Many of the studies described in this thesis are of a collaborative nature, and many were performed as part of MalariaGEN Consortium. Below is a summary of the contributions of other scientists to the work described in this thesis.

Chapters 3 and 4

The candidate gene case-control data used for testing the association and epistasis of malaria candidate genes were generated as part of the MalariaGEN Consortial Project 1 (CP1). The TW lab team performed the DNA extraction at KEMRI-Wellcome Trust Research Programme in Kilifi. Sample processing, quality control (QC) and Sequenom genotyping was carried out by the MalariaGEN Resource Centre at the Wellcome Trust Centre for Human Genetics in Oxford UK. The α^{+}thalassaemia typing was both performed by the TW lab-team and the MalariaGEN Resource Centre (MRC).

Chapter 5 and 6

The genome-wide association study (GWAS) used in these two chapters was part of the GWAS conducted by the MalariaGEN consortium. The collection of all the samples and phenotype data were done by the TW group while sample selection was performed by the MalariaGEN consortium. Sequencing was carried out at the Wellcome Trust Sanger Institute, and sequence data was processed and quality controlled by the MRC. Variant calling, imputation,

stringent sample and variant QC and association analysis were carried out by Gavin Band, Si Quang Le, Chris Spencer, Geraldine Clark and Kirk Rockett. Quang helped in selecting the SNPs used in the characterisation of the haplotype structure in Chapter 5, while Gavin assisted in enhancing the SNPepistasis method for detecting epistatic interactions in the GWAS study by incorporating a wrapper script and also he wrote custom software to split the GWAS data into chucks. He is also supported in running the genome scans.

Contents

List of Figures

List of Tables

Abbreviations

BCS	Blantyre Coma Score
CI	Confidence Interval
CM	Cerebral Malaria
CR1	Complement Receptor 1
COD	Cause of Death
DNA	Deoxyribose Nucleic Acid
EIR	Entomological Innoculation Rate
G6PD	Glucose-6-Phosphate Dehydrogenase
GWAS	Genome-Wide Association Study
GWIS	Genome-Wide Interaction Scan
Hb	Haemoglobin
HBA	Haemoglobin Alpha
HBB	Haemoglobin Beta
HbS	Haemoglobin S
HWE	Hardy-Weinberg equilibrium
HP	Haptoglobin
ICAM-1	Intercellular Adhesion Molecule-1
IFN	Interferon
IL	Interleukin
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health
iRBC	Infected Red Blood Cell
KEMRI	Kenya Medical Research Institute
KCH	Kilifi County Hospital

KHDSS	Kilifi Health Demographic Surveillance System
KM^2	Square Kilometre
KNBS	Kenya National Bureau of Statistics
KWTRP	KEMRI/Wellcome Trust Research Programme
LD	Linkage Disequilibrium
LRT	Likelihood Ratio Test
LTA	Lymphotoxin-Alpha
MAF	Minor Allele Frequency
LIMS	Laboratory Information Management Systems
MalariaGEN	Malaria Genomic Epidemiology Network
Mb	Megabase
OR	Odds Ratio
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PEP	Primer Extension Pre-amplification
Q-Q Plot	Quantile-Quantile plot
RBC	Red blood Cell
RD	Respiratory Distress
RFLP	Restriction Fragment Length Polymorphism
SMA	Severe Malaria Anaemia
SM	Severe malaria
SNP	Single Nucleotide Polymorphism
SSA	sub-Saharan Africa
TNF-α:	Tumour Necrosis Factor-Alpha
WHO	World Health Organization
WTCHG	Wellcome Trust Centre for Human Genetics

Chapter 1

Introduction

Abstract

In this chapter, I will describe the background to my thesis. I will start by describing the world malaria situation with an emphasis on Africa and in particular Kenya. Subsequently, I will review the life cycle of the malaria parasite and the main clinical manifestations of malaria. I will outline some of genetic principles that underlie my thesis before reviewing the literature on some of the host genetic factors that have been shown to play a significant role in determining an individual's susceptibility to malaria that are of particular interest in the context of my thesis. Much of what follows will be about multi-locus effects and explain why considering potential interactions between malaria candidate genes or "epistasis" might be important if we are to gain a better understanding of their roles in malaria pathogenesis. Finally, I will discuss the statistical methods that are commonly used for identifying both disease-association genes and gene-gene interactions and enumerate the primary questions that underlie the rest of my thesis.

1.1 Epidemiology of Malaria

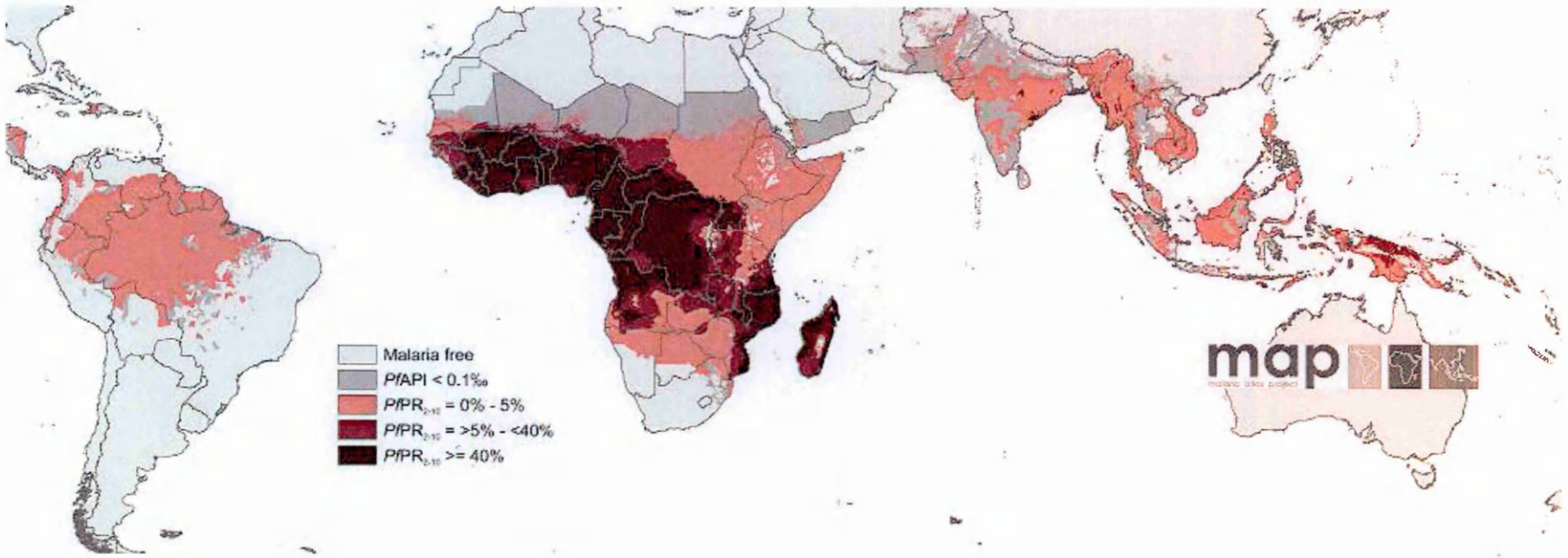
Malaria is a devastating parasitic infection of major public health concern in many parts of the tropical world [1]. The World Malaria Report, 2014 [2] estimated that around 200 million cases of clinical malaria occurred in 2013, of which >80% occurred in the Africa region and >90% were caused by *Plasmodium falciparum,* [2] and that the number of deaths due to malaria was 584,000 [2] of which almost 80% occurred in children aged under 5 years.

Among the 35 countries that accounted for almost 98% of the total malaria deaths globally, 30 were located in sub-Saharan Africa (SSA). These were: Nigeria, Democratic Republic of Congo, Uganda, Ethiopia, Tanzania, Sudan, Niger, Kenya, Burkina Faso, Ghana, Mali, Cameroon, Angola, Côte d'Ivoire, Mozambique, Chad, Guinea, Zambia, Malawi, Benin, Senegal, Sierra Leone, Burundi, Togo, Liberia, Rwanda, Congo, Central African Republic, Somalia, and Guinea-Bissau [3].

Pregnant women in malaria-endemic countries and non-immune travellers are also at high risk. Worldwide, great and varied efforts are being made to learn more about this disease and how to control it. Management strategies today include education, the development of vaccines and chemotherapeutic agents and

vector control using approaches that include indoor residual spraying with insecticides and the provision of insecticide treated bed nets. Although these efforts have met with some success, malaria still remains a leading cause of morbidity and mortality, particularly in SSA [2], and parts of Asia and Latin America (Figure 1.1).

Figure 1.1. Global distribution of countries and areas at risk of *P. falciparum* malaria in 2010.



The map shows endemicity predictions based on *P. falciparum* Parasite Rates (PfPR). Predictions were categorized as low risk PfPR 2-10≤5% light red; intermediate risk PfPR 2-10>5% to <40%, medium red; and high risk PfPR 2-10 ≥40%, dark red. The rest of the land area was defined as unstable risk (medium grey areas, where *Pf* API<0.1 per 1,000 pa) or no risk (light grey). Adapted from Gething *et al.* 2011 [4].

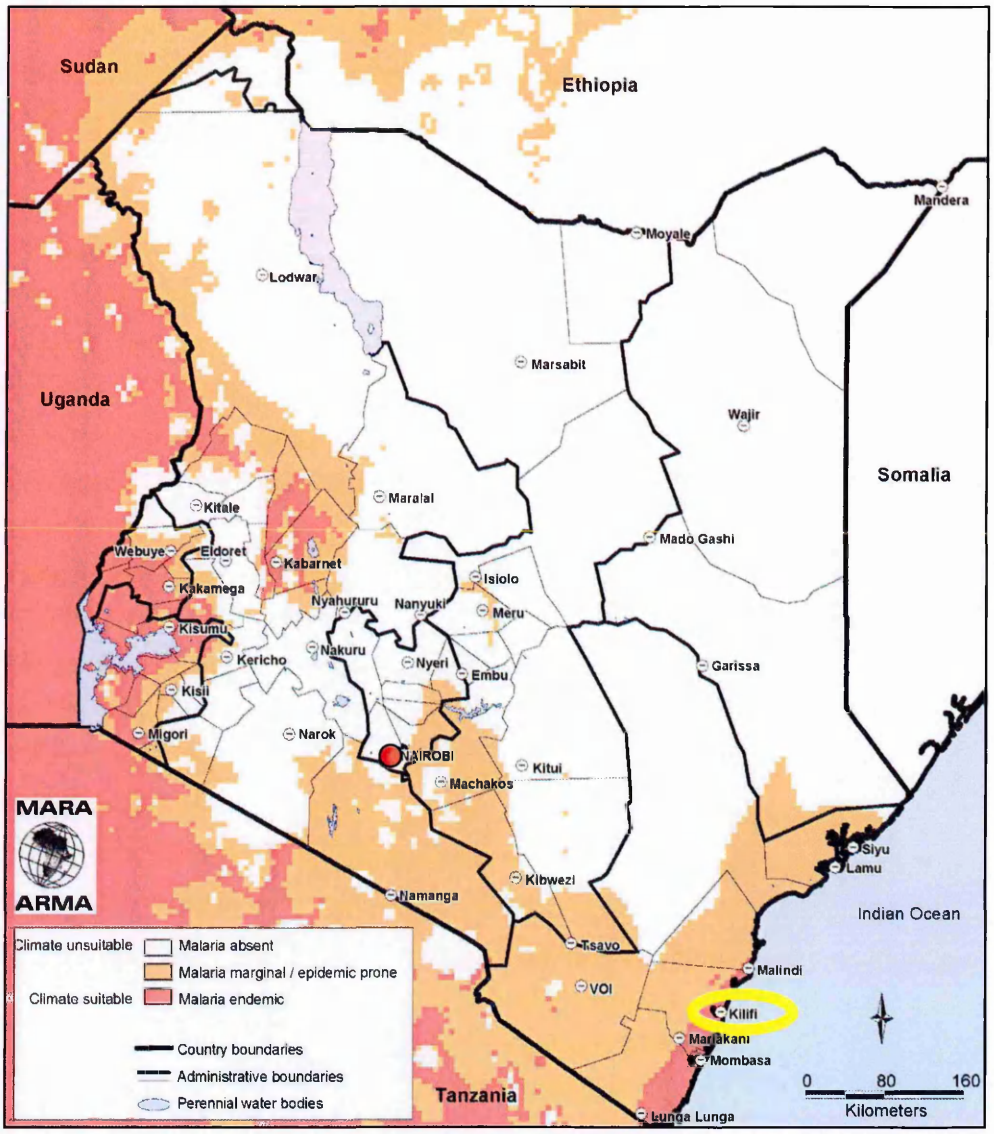
1.1.1 The malaria situation in Kenya

Kenya is situated in East Africa and is bordered by Ethiopia to the north, Sudan to the northwest, Somalia to the east, Tanzania to the south and Uganda to the west (Figure 1.2). Kenya covers a total land mass of 582,646 km^2, has a current population of 38.6 million [5] and is divided into 47 administrative counties. The national life expectancy is 61 years, 43% of the population are under the 15 years and 4% aged 65 years and above [5]. The maternal mortality ratio is 360 deaths per 100,000 live births, the infant mortality rate is 48 deaths per 1000 live births, and under 5 mortality is 73 per 1000 live births [6]. These rates are high in comparison to countries in the Western world: for example, in England and Wales the maternal mortality ratio was 12 per 100,000 live births, the infant mortality rate was 4 per 1,000 live births and the under 6 mortality was 4 per 1000 live births in 2012 [7]. The major causes of death in children less than 5 years of age in Kenya are neonatal causes, pneumonia, malaria and diarrhoeal diseases [6].

Kenya has four epidemiological zones for malaria (Figure 1.2); endemic areas along the shores of Lake Victoria and the coastal region where malaria transmission is seasonal, with peaks from June to August and again in late November; highland-epidemic-prone areas which are highly populated; epidemic

prone areas in the arid and semi-arid lowlands which are sparsely populated; and low-risk or malaria-free areas in the highlands. Transmission in the epidemic prone areas is highest from April through June each year [8]. About 70% of the population of Kenya is at risk from malaria [9] which is responsible for about 30% of out-patient visits (totalling more than eight million out-patient treatments each year), and 19% of all hospital admissions [10]. Approximately, 14,000 children are admitted annually to a hospital due to malaria. It is also estimated that 6,000 pregnant women annually suffer from malaria-associated anaemia and that approximately 4,000 babies are born annually with low birth weight due to maternal anaemia. Malaria imposes a significant economic burden in Kenya, it is estimated that in 2009, 170 million working days were lost due to malaria illness [10]. Overall malaria is falling in Kenya, and this could be partly due to local initiatives. However, several studies show an increase in malaria within the epidemic-prone and low transmission zones [11]. One of the main study/research areas for malaria and other diseases lies on the coast north of Mombasa in Kilifi County (marked by the yellow circle in Figure 1.2) which falls within the malaria endemic zone. The malaria studies described in this thesis were undertaken at the KEMRI-Wellcome Trust Research Programme in Kilifi and are discussed in more detail in Chapter 2.

Figure 1.2. Distribution of endemic malaria in Kenya.



The map is showing Kenya and the surrounding countries with international borders, the national capital Nairobi (the red dot) and its 47 administrative counties. The Kilifi County, where the sample collection took place in this thesis is marked by the yellow circle. The map also illustrates the malaria endemic areas within the country stratified by climatic suitability. Adapted from the Mapping Malaria in Africa project (www.mara.org.za).

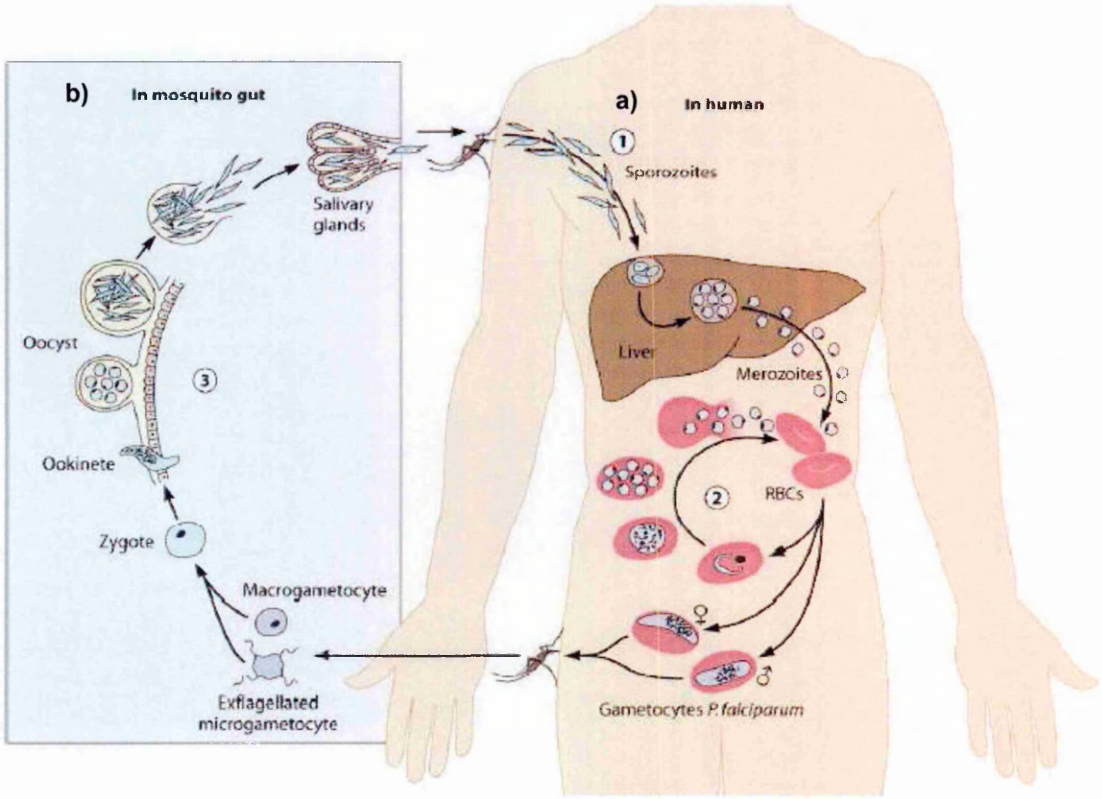
1.1.2 Malaria parasites and the Malaria Life Cycle

Human malaria is caused by protozoan parasites of the genus *Plasmodium.* Five species have been recognized to cause malaria in humans: *P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi.* *P. falciparum* is the most important in SSA, accounting for the majority of deaths [12], is the only species of relevance in Kilifi and is therefore the focus of my thesis.

To complete its life cycle the malaria parasite requires two hosts, the definitive host, in which the parasite undergoes sexual development (a female mosquito), and an intermediate host; anopheles mosquitoes and humans in the case of human malaria. Figure 1.3 illustrates the life cycle of *Plasmodium* species using *P. falciparum* as an example. The infection starts when a female mosquito begins to feed on a human host. Sporozoites are injected from the salivary glands into the dermis of the human from where they make their way into the blood vessels. The sporozoites then travel to the liver where they invade hepatocytes and multiply [13]. After about 10 days the hepatocytes rupture, releasing thousands of merozoites into the bloodstream. The merozoites rapidly invade the red blood cells (RBCs) initiating an asexual growth phase in the blood that leads to the clinical symptoms seen in malaria.

Over a cycle of 48 hours, the parasite develops within the RBC from the ring to trophozoite stage before replicating asexually to form schizonts composed of twelve to twenty-four merozoites. Lysis of the red cell releases the merozoites which invade fresh RBCs to start the next cycle [13]. A small sub-set of the invading merozoites develop into gametocytes (male and female), which are taken up by feeding Anopheline mosquitoes. Gametocytes combine sexually in the mosquito mid-gut before developing sequentially into zygotes, ookinetes and then oocysts in the gut wall. These burst into the mosquito haemocoel releasing sporozoites that migrate to the salivary glands to await inoculation into the human host at the next blood meal [13].

Figure 1.3. The Life cycle of *P. falciparum.*



a) Life stages within the human host. b) Life stages within a female mosquito of the genus *Anopheles.* Adapted from Bousema and Drakely 2011 [13], and published with permission from American Society for Microbiology).

1.1.3 The clinical manifestations of malaria

The clinical outcomes of *P. falciparum* infection range from asymptomatic infection to severe disease and death. In many malaria-endemic regions of Africa, the majority of individuals carry *P. falciparum* parasites during the high malaria transmission season. While most individuals are asymptomatic or display only mild symptoms of malaria disease such as fever, headache, nausea, vomiting and cough, 1-2% of infections progress to severe and potentially fatal complications [14]. The diagnosis of malaria is based on clinical criteria supplemented by the detection of parasites in the blood [15].

1.1.3.1 Severe malaria

The WHO defines severe malaria (SM) as one or more of the following symptoms in a patient with *P. falciparum* malaria: prostration (a Blantyre coma score (BCS) of 3 or 4), cerebral malaria (CM) (a BCS of <3), multiple seizures (two or more seizures within 24 hours prior to admission), severe malaria anaemia (SMA) (haemoglobin <5g/dl or a haematocrit <15% in association with a parasitaemia >10,000/μl of blood) or respiratory distress (RD) (abnormally deep breathing).

The most common clinical manifestations of SM in childhood are SMA, CM and RD [16]. SM is almost exclusively caused by *P. falciparum* and is a major cause

of childhood hospital admission and death in endemic areas [17]. The frequency, pattern and distribution of severe forms of *P. falciparum* malaria differ depending on the transmission areas [18, 19]. A patient with SM symptoms has a high chance of dying, but this depends on a range of factors that include host genetic makeup, background immunity, age and access to clinical care [20].

1.1.3.2 Severe malarial anaemia

The greatest burden of SMA is carried by young children and pregnant women [21]. Mortality from anaemia is about 1% when occurring alone but increases rapidly (up to 35%) when occurring in combination with other severe complications including CM or RD as shown in Figure 1.4 [22]. Apart from malaria, other superimposed conditions may serve to increase the risk to SMA, including the presence of hookworm infections or iron deficiency [23]. In SSA, strong evidence support an age-dependent pattern of SM disease, with SMA being more common in infants and young children under three years and CM in older children [21, 24].

1.1.3.3 Cerebral malaria

Cerebral malaria is one of the most severe complications of *P. falciparum* infection.. Mostly, patients present with fever, headache, and delirium before progressing to an acute febrile stupor, followed by coma [25]. It is a major cause

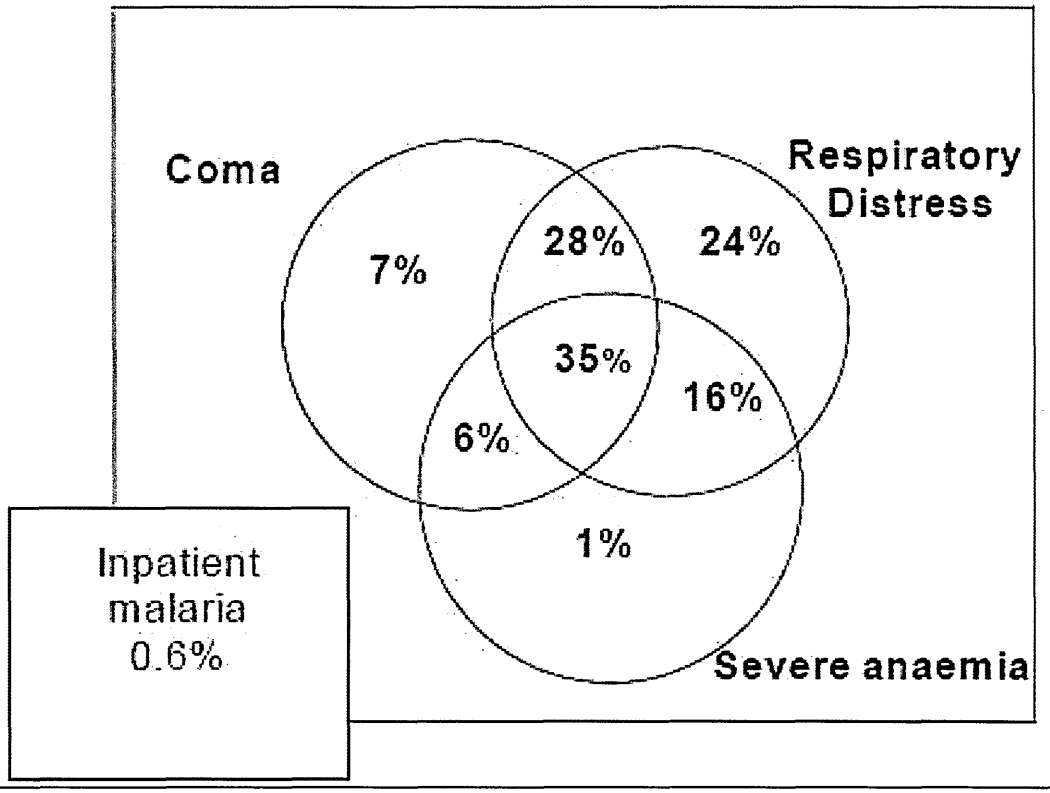
of mortality in *P. falciparum* infection. Approximately 1% of infections progress to CM. The majority of cases occur in SSA, and approximately 15-30% of individuals who develop CM die [26].

1.1.3.4 Respiratory distress

The pathogenesis of RD in patients with malaria infection is not well understood; it is most commonly thought to be a clinical manifestation of metabolic acidosis and is associated with a mortality of approximately 24% [22]. Mortality increases rapidly (to >30%) when RD is found in combination with CM [22].

Since the purpose of genetic epidemiology in malaria is to unravel polymorphisms in molecular pathways that may eventually translate into new ways of preventing or treating the disease, SM and its sub-phenotypes (CM, SMA and RD) are the phenotypes of choice because they increase chances of detecting differences for most studies and the study undertaken for this thesis is no exception.

Figure 1.4. Venn diagram comparing the mortality of the three major sub-phenotypes of severe malaria to other children who were hospitalised with *P. falciparum* malaria.



The inpatient malaria box represents other children hospitalised with *P. falciparum* malaria. (Adapted from Maitland and Marsh 2004 [27], and published with permission from ScienceDirect).

1.2 The human genome

A "genome" is an organism's complete set of deoxyribonucleic acid (DNA); hereditary information that is passed from generation to generation. DNA molecules are made of a pair of twisted strands. Each strand consists of a sugar-phosphate backbone with units called "nucleotide bases" hanging off it. There are 4 forms of bases: adenine (A), thymine (T), cytosine (C) and guanine (G), and the human genome can be considered as a very long sequence consisting of approximately 2.9 billion nucleotides, each denoted by the letter of its names, e.g. ATCCGA. The two strands are held together by primarily hydrogen bonding between the bases. Adenine is always paired with T and C is always paired with G, actually both strands contain the same information. These paired nucleotides are called "base pairs".

The main function of DNA is to act as a template for the production of proteins, which are functional units participating in most cellular processes. In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called "chromosomes". The human genome comprises 23 pairs of chromosomes as 22 pairs of autosomes and one pair of sex chromosome (a female has two X-chromosomes whereas a male has one X and one Y sex chromosome): 46 chromosomes in total. Because the autosomal chromosomes are arranged in

pairs, they are called "homologous" chromosome pairs. In each pair, one of the chromosomes is inherited from the mother and the other from the father. A "sex chromosome" is a type of chromosome that participates in sex determination. A specific location along a given chromosome, or a specific position in the genome, is called a "locus". The term "gene" is sometimes defined as a segment of DNA within a chromosome that specifies the amino acid sequence for a peptide/protein.

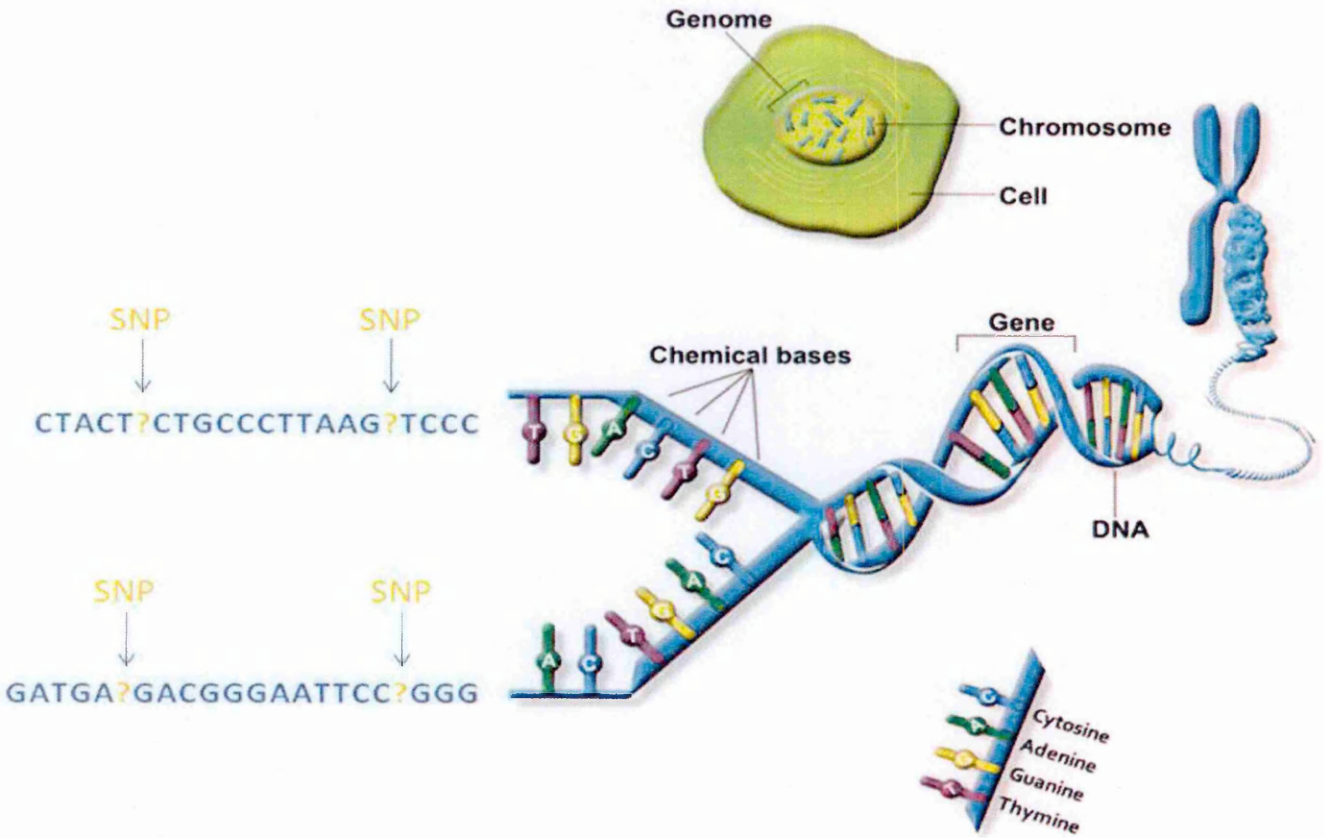
The majority of DNA is similar between the genomes of all humans. Nevertheless, due to changes in the nucleotide sequences of DNA molecules that have occurred during population history, there is variability at millions of locations in the genome. There are several types of such variation, commonly called "polymorphisms". Of these, the most common are "single nucleotide polymorphisms" (SNPs), which are characterised by single base changes in DNA sequence. Figure 1.5 is an illustration of DNA, gene and SNP in a chromosome in the genome. The alternative DNA sequences at a locus, or different "versions" of the same gene, are known as "alleles". A "reference" allele for a given locus refers to the wild type or the ancestral while a "derived" allele is an allele that arises in the evolution due to a mutation. For a given polymorphism, the combination of alleles from two chromosomes determines the "genotype" of the

individual. The "homozygous" genotype is when two alleles are identical (i.e. AA), and "heterozygous" genotype is when the alleles are different (i.e. AT).

A combination of alleles at different loci on a chromosome that are inherited together is referred to as a "haplotype". Linkage disequilibrium (LD) is the correlation or non-random association of alleles at adjacent loci. When a particular allele at one locus is found together on the same chromosome with a specific allele at a second locus more often than expected if the loci were segregating independently in a population, the loci are in LD.

Many types of polymorphism occur but the most common after SNPs are "insertion/deletion" of one or more base pairs from the sequence, "inversions", arrangements in which order of the DNA sequence is reversed in a specific chromosomal region, among others "copy number variants", segments of DNA which are present at a different copy number compared to the reference genome. Further details regarding these concepts and terminologies can be found in review papers such as Olson *et al.* [28].

Figure 1.5. An illustration of the structure of DNA, gene and SNPs in a chromosome in the genome.



The arrows indicate the position of each SNP on a chromosome. (Adapted from Coriell Institute for Medical Research [29]).

1.3 Host genetics factors and malaria

Malaria has been a major cause of mortality since ancient times [30] and, as a result, it has been a powerful selective driving force in humans as both parasite and host have evolved to increase biological fitness through natural selection. The variations that have occurred in the human genome over multiple generations to yield so-called 'malaria protective genes' are not only of scientific importance but are also biologically important due to their impact on health.

While numerous candidate genes have been screened for their involvement in malaria disease severity [31], most have been investigated in small studies.
A longitudinal survey of twin children in The Gambia [32] was one of the first genetic studies to address the question of genetic determinants of susceptibility to clinical malaria. In this study, the authors compared the incidence of clinical malaria in pairs of twin children within families who shared the same environmental exposure. The monozygotic twins (children derived from a single zygote who are therefore genetically identical) were both found more likely to have a malaria fever attack than dizygotic twins (children derived from two zygotes), signifying a role of host genetic factors on disease burden.

The contribution of genetic and non-genetic factors towards SM was also examined in a Sri Lankan population, which was one of the first genetic studies to quantify the contribution of genetic factors to susceptibility to clinical malaria [33]. The heritability was approximately 15% for the incidence of both *P. falciparum* asymptomatic and symptomatic infections, and around 10% for the intensity of clinical symptoms. Through a similar study, it has been estimated that 34% of the total variation in both uncomplicated and hospital-admitted malaria in two cohorts of Kenyan children was explained by additively acting host genes [34]. However, sickle cell trait (HbAS), the most prominent of the malaria-protective polymorphisms identified to date only counted for around 2% of this variation [34], suggesting that more malaria protective genes remain to be discovered. In a subsequent study conducted in Thailand similar observations have been made, where the contribution of genetic factors to the variability in the incidence of uncomplicated *P. falciparum* and *P. vivax* malaria infections in a large cohort of mixed ages was found to be 10% and 19% respectively [35].

A wide range of other candidate genes have been demonstrated to confer protection against SM as summarised in Table 1.1. For example, genetic variants of β-globin: HbS [36], HbE [37], HbC [38]; α-globin [39-41]; enzymopathies (e.g. glucose-6-phosphate dehydrogenase deficiency) [42, 43]; structural variants of red

cell membrane proteins (e.g. band 3 protein (AE1) [44], *ATP2B4* [45], *HLA*) [46] and several cytokine loci (e.g. Tumor Necrosis Factor-alpha [47-49], Interleukin-12 [50], Interferon-alpha receptor- 1 [51], Interleukin-4 [52]). Nevertheless, the majority of these studies have been small, and some of them have yielded conflicting results.

With recent advances in genotyping technologies and analyses using advanced molecular biology software, it is likely that the already large panel of host genetic factors will expand and thereby also increase the likelihood of identifying true associations. In this thesis, I seek to confirm some of the above reported associations using a large case-control study of SM conducted in Kenya. The results of this study are presented in Chapter 3. In the sections that follow, I will discuss some of the haemoglobinopathies (disorders of the haemoglobin structure or production) that are of particular interest in the context of this thesis.

Table 1.1. Genetic polymorphisms implicated in susceptibility/resistance to *P. falciparum* malaria in Africans studies.

Gene Name	Symbol	Phenotype	Result	Population	Reference
ABO blood group	*ABO*	Malaria severity, protection from higher parasitaemia in uncomplicated malaria, severe infection	+	Gabon, The Gambia, Mali, Kenya, South Africa (Malawi)	[53-56]
		Malaria incidence and onset, multiplicity of *P. falciparum* infection	−	Mali, Senegal	[57, 58]
Allograft inflammatory factor 1	*AIF1*	SM	+	The Gambia	[59]
Apolipoprotein E	*APOE*	Age of infection, uncomplicated malaria and SM	−/(+)	The Gambia, Ghana	[60]
α^{+}thalassaemia		Anaemia in asymptomatic malaria, multiplicity of *P. falciparum* infection, uncomplicated malaria, SM and SMA	+	Ghana, Kenya, Nigeria, Senegal, Tanzania	[61-66]
		Low density parasitaemia, malaria incidence and onset, malaria severity, *P. falciparum* infection	−	Gabon, Mali, Nigeria, Tanzania	[55, 57, 63, 67]
CD36 molecule	*CD36*	SM	+	The Gambia, Kenya	[68-71]
			−	The Gambia, Ghana, Kenya, Malawi, Nigeria	[53, 54]
CD40 ligand	*CD40LG*	SM	+	The Gambia	[72]
Chr2q37.1		SM	+	The Gambia	[73]
Chr5p15 region		Number of clinical malaria attacks	(+)	Senegal	[55]
Chr5q31 region		Asymptomatic parasite density	+	Senegal	[55]
Chr5q31-q33 region		Blood infection level	+	Burkina Faso, Cameroon	[56, 58, 59]
Chr7p12.2		SM	+	The Gambia	[73]
Chr7q32.3		SM	+	The Gambia	[73]

Table 1.1 cont.

Gene Name	Symbol	Phenotype	Result	Population	Reference
Chr11p15.4 (coding region of HBB)		SM	+	The Gambia	[73]
Chr12q21 region		Maximum parasite density during asymptomatic infection	(+)	Senegal	[55]
Chr13q13 region		Number of clinical malaria attacks	(+)	Senegal	[55]
Chr17p13.1		SM	+	The Gambia	[73]
Chr21q22.11 region		SM	+	The Gambia, Kenya	[61, 73]
Complement component receptor 1	*CR1*	SM	+	Kenya	[63]
			−	The Gambia, Mali	[64, 65, 67, 68]
Duffy blood group gene		*P. vivax* infection	+	Kenya	[69]
Fc fragment of IgG, low affinity IIa, receptor (CD32)	*FCGR2A*	High-density parasitaemia, SMA, placental and SM	+	The Gambia, Ghana, Kenya, Sudan	[69, 70, 74-77]
		SM	−	Kenya	[78]
Fms-related tyrosine kinase 1	*FLT1*	Malaria resistance *in utero*	+	Tanzania	[79]
Fucosyltransferase 9	*FUT9*	Placental malaria	+	Mozambique	[80]
Glucose-6-phosphate dehydrogenase	*G6PD*	Asymptomatic and uncomplicated malaria, malaria incidence in children, multiplicity of *P. Falciparum* infection	+	Gabon, The Gambia, Mali, Senegal, Uganda	[42, 43, 81-87]
		Low density parasitaemia, malaria incidence and onset, malaria severity	−	Gabon, Mali, Tanzania,	[57, 88]
Glutathione S-transferase mu 1	*GSTM1*	Mild and SM	+	Cameroon	[89]
		SMA	−	Tanzania	[90]
Glutathione S-transferase pi 1	*GSTP1*	SMA, uncomplicated and mild malaria	+	Cameroon, Tanzania	[89, 90]
Glutathione S-transferase theta 1	*GSTT1*	SMA, uncomplicated, mild and SM	−	Cameroon, Tanzania	[89, 90]
Glycophorin A	*GYPA*	Erythrocyte invasion	−	The Gambia	[91]

Table 1.1 cont.

Gene Name	Symbol	Phenotype	Result	Population	Reference
GNAS complex locus	*GNAS*	Erythrocyte invasion	+	Ghana, The Gambia, Kenya, Malawi	[92]
Haptoglobin	*HP*	Clinical and SM, parasitaemia, placental infection by *P. Falciparum*	+/-	Cameroon, Ghana, The Gambia, Kenya	[93-96]
		SM and uncomplicated malaria	−	Gabon, The Gambia, Ghana	[83, 97, 98]
Haemoglobin, beta	*HBB*	Asymptomatic, mild, uncomplicated and SM, malaria incidence in children, malaria incidence and onset, malaria severity, parasite density, symptomatic parasitaemia, SMA	+	Burkina Faso, Gabon, The Gambia, Ghana, Kenya, Liberia, Malawi, Mali, Senegal, Tanzania, Uganda	[38, 42, 57, 62, 78, 81, 88, 99-102]
		Multiplicity of *P. Falciparum* infection, SM and symptomatic malaria	−	Kenya, Mali, Senegal, Uganda	[67, 85, 87, 100]
Haemoglobin S and α^+thalassaemia combination		SM and α^+thalassaemia	−	Kenya	[103]
Intercellular adhesion molecule 1 (CD54)	*ICAM1*	Asymptomatic, mild, SM and CM	+	Gabon, Kenya, Nigeria	[53, 104, 105]
		SM and CM	−	The Gambia, Kenya, Senegal, South Africa, Malawi	[64, 78, 106]
HLA-B associated transcript 2	*BAT2*	SM	+	The Gambia	[107]
Interferon, alpha 2	*IFNA2*	SM	−	Gabon	[108]
Interferon, gamma	*IFNG*	SM	+	The Gambia, Mali	[109]
Interferon (alpha, beta and omega) receptor 1	*IFNAR1*	SM	+	The Gambia, Kenya	[43, 56]
		SM	−	Kenya	[58]
Interferon receptor 2	*IFNAR2*	SM	+	The Gambia	[56]
Interferon gamma receptor 1	*IFNGR1*	SM	+	The Gambia	[110]
Interferon gamma receptor 2	*IFNGR2*	SM	−	The Gambia	[56]
Interferon regulatory factor 1	*IRF1*	SM	+	Burkina Faso	[87]
		SM,CM and SMA	−	The Gambia, Kenya, Malawi	[87]

Table 1.1 cont.

Gene Name	Symbol	Phenotype	Result	Population	Reference
Interleukin 1, alpha	*IL1A*	SM and uncomplicated malaria	+	The Gambia	[111]
		Uncomplicated and CM, severe anaemia	–	Ghana	[112]
Interleukin 1, beta	*IL1B*	IL1β production, parasitaemia density in uncomplicated malaria, SMA and SM	+	The Gambia, Ghana, Kenya	[111-113]
		Asymptomatic parasite density, CM, clinical and uncomplicated malaria, severe anaemia	–	Ghana, Tanzania	[112, 114]
Interleukin 1 receptor antagonist	*IL1RN*	Asymptomatic parasite density, clinical and SM	–	The Gambia, Tanzania	[64, 114]
Interleukin 4	*IL4*	Level of *P. Falciparum* specific antibodies, SM	+	Burkina Faso, Ghana, Mali	[87, 109, 112, 115]
		Asymptomatic parasite density, clinical malaria	–	Burkina Faso, Tanzania	[114, 116]
Interleukin 10	*IL10*	IL10 production, SMA	+	Kenya	[113]
		Asymptomatic parasite density, clinical and SM	–	The Gambia, Tanzania	[114, 117]
Interleukin 10 receptor, beta	*IL10RB*	SM	–	The Gambia	[97]
Interleukin 12B	*IL12B*	CM	+	Mali, Tanzania	[50, 118]
		Parasitaemia	–	Burkina Faso	[119]
Interleukin 13	*IL13*	Asymptomatic parasite density, clinical malaria	–	Tanzania	[114]
Interleukin 22	*IL22*	SM	(+)	The Gambia	[120]
Interleukin 26	*IL26*	SM	–	The Gambia	[120]
Locus 10p15 (*P. falciparum* Fever Episodes Quantitative Trait Locus 1)	*PFFE1*	Uncomplicated malaria	(+)	Ghana	[121]

Table 1.1 cont.

Gene Name	Symbol	Phenotype	Result	Population	Reference
Macrophage migration inhibitory factor	*MIF*	High-density parasitaemia	+	Kenya	[122]
		SMA	−	Kenya	[122]
Lymphotoxin alpha	*LTA*	Asymptomatic parasite density, parasitaemia, SM	+	Burkina Faso, Malawi, Tanzania	[42, 114, 119]
		Clinical and SM	−	The Gambia, Kenya, Tanzania	[42, 107, 114]
Macrophage migration inhibitory factor	*MIF*	SMA	+	Kenya	[122]
Macrophage stimulating 1	*MST1*	Asymptomatic and SM	+	Nigeria	[53]
Major histocompatibility complex	*HLA*	Asymptomatic parasite density, malaria, SM	+	Gabon, The Gambia, Ghana, Tanzania	[114, 123, 124]
		Malaria	(+)	The Gambia	[125]
Major histocompatibility complex	*HLA*	Clinical malaria	−	Tanzania	[114]
Mannose-binding lectin 2	*MBL2*	Asymptomatic, placental, SM and uncomplicated malaria	−	Cameroon, Gabon, The Gambia	[64, 83, 84]
Natural cytotoxicity triggering receptor 3	*NCR3*	Uncomplicated malaria	+	Burkina Faso	[126]
Nitric oxide synthase 2, inducible	*NOS2*	SM, SMA	+	Gabon, The Gambia, Ghana, Kenya, Tanzania	[127-130]
		Asymptomatic, symptomatic, uncomplicated and SM	−	Gabon, Kenya, Tanzania, Uganda	[78, 83, 84, 131]

Table 1.1 cont.

Gene Name	Symbol	Phenotype	Result	Population	Reference
Platelet/endothelial cell adhesion molecule (CD31 antigen)	*PECAM1*	SM	–	Kenya, Nigeria	[53, 132]
Solute carrier family 4, anion exchanger, member 1 (anion-exchange protein 1, AE1)	*SLC4A1*	SMA and fatality	+	Ghana	[133]
T-cell antigen receptor, beta subunit	*TCRB*	Asymptomatic parasite density	+	Tanzania	[114]
		Clinical and SM	–	The Gambia, Tanzania	[114, 134]
Toll-interleukin 1 receptor (TIR) domain containing adaptor protein	*TIRAP*	SM and general malaria	+/(+)	The Gambia, Kenya	[61, 135]
		Placental malarial infection	–	Ghana	[136]
Toll-like receptor 9	*TLR9*	SM	–	The Gambia, Malawi	[136]
Transporter 1, ATP-binding cassette, sub-family B	*TAP1*	SM and uncomplicated malaria	+	Gabon	[126]
Triosephosphate isomerase 1	*TPI1*	Malaria infection	–	Angola, Mozambique, Sao Tomé e Prìncipe	[137]
Tumor necrosis factor	*TNF*	Asymptomatic, clinical, SM and uncomplicated malaria, malarial fever, malaria morbidity and mortality, SMA	+	Burkina Faso, Gabon, The Gambia, Kenya, Tanzania	[42, 107, 114]
		Asymptomatic parasite density, symptomatic, uncomplicated and SM	–	Gabon, Kenya, Malawi, Mali, Tanzania, Uganda	[81, 83, 99, 109]

+ indicates a genetic association or linkage was reported ($p \leq 0.05$), (+) indicates a suggestive linkage or association was reported, – indicates no genetic association was detected, +/- indicates different studies have contradictory results. SM=severe malaria; CM=cerebral malaria; SMA=severe malarial anaemia; RD=respiratory distress.

1.3.1 Sickle haemoglobin (HbS)

The HbS variant is caused by a single base change in codon 6 of the β-globin gene (*HBB*) that leads to the substitution of valine for glutamate within the β-globin protein. The resulting variant form of haemoglobin, HbS, tends to polymerise under conditions of low oxygen tension, becoming insoluble and causing distortion of erythrocytes, which adopt a rigid 'sickle' shape [138]. The carrier form (heterozygote - one normal (HbA) and one abnormal (HbS) copy of the *HBB* gene) of HbS referred to as sickle cell trait (HbAS) was among the first genetic conditions to be associated with malaria protection more than 60 years ago [17, 139, 140]. I will further elaborate in detail this polymorphism and its relationship with SM in subsequent Chapters.

1.3.2 Thalassaemia

The thalassaemias are inherited conditions of haemoglobin that are manifested by decreased or zero production of the α– or β-globin proteins that are required for the formation of haemoglobin. Unlike HbS, the thalassaemias do not affect the structure of the globin protein itself, but according to its form, leads to deficient or absent synthesis of α– or β-globin chains. Alpha-thalassaemia occurs when one or more copies of the paired α-globin genes on chromosome 16 are missing, most commonly due to the inheritance of deletions within *HBA* [38].

Alpha-thalassaemia is found at high frequencies in many parts of Africa, South and South-east Asia and the Pacific, as well as in African descendants in South and Central America. In some countries (most notably Nepal and Papua New Guinea) the prevalence of α^+thalassaemia approaches 80%, while in SSA prevalence does not usually exceed 50% [141]. Beta-thalassaemia, on the other hand, results from defects in the *HBB* gene on chromosome 11 that encode the production of the beta globin protein [120]. Beta-thalassaemia is prevalent in the Mediterranean region including the coastal areas of Turkey, Greece and Italy [142]. In our previous work in Kilifi, we found no evidence of β-thalassaemia in our population [143] therefore, in this thesis I will only focus on α^+thalassaemia.

1.3.3 The ABO Blood group system

The ABO blood type was first discovered in 1900 by Karl Landsteiner [144]. The ABO blood group system is the most important RBC polymorphism due to its significant role in successful blood transfusion and organ transplantation [145]. The system is split into 3 groups called A, B and O. These groups are defined on the basis of the composition of the oligosaccharides bound to the H-antigen on the red cell surface. The O blood group lacks these oligosaccharides due to a non-functional transferase enzyme resulting from a frame-shift mutation, while blood groups A and B are the result of genetic mutation in the same transferase

enzyme affecting its sugar specificity[146]. A is the ancestral type on which a polymorphism arose to create B. Then later a further mutation arose on the A haplotype causing a frame-shift effectively knocking-out the ABO locus, this is the O type [147]. There are many other mutations in the ABO gene leading to sub–type of each of the 3 main haplotypes. Since the discovery of the ABO gene, numerous association studies have investigated the relationships between ABO blood group antigens and susceptibility, resistance, or severity of *P. falciparum* malaria disease. Some have found associations between the ABO system and severity of the disease, while others have reported an absence of an association between malaria disease and the ABO blood groups antigens [145, 148]. Chapter 3 of this thesis describes the results of my findings.

1.3.4 Glucose-6-phosphate dehydrogenase (G6PD) deficiency

G6PD is a housekeeping enzyme that functions as an antioxidant, and thus protects the cells from oxidative damage. The *G6PD* gene is located on the X chromosome. Many allelic variants are found at this locus, two of which can reach frequencies of up to 20% in populations from Africa and the Mediterranean [149]. These 2 alleles are known as G6PD c.202G>A (rs1050828) and G6PD c.376A>G (rs1050829) and together fall into the A- WHO classification scheme (the A being an electrophoretic mobility shift due to 376G

and the minus being a decrease in enzyme activity driven by 202A). This pair of alleles (as a haplotype) is the cause of the commonest G6PD deficiency in SSA exhibiting about 12% of normal enzyme activity (or B type in the same WHO classification) [150]. Since this disorder is sex-linked (females have 2 copies of the gene and males one copy (hemizygous)), it affects almost exclusively males; however, homozygous females are also affected and heterozygous females can also suffer from haemolytic anaemia induced by G6PD deficiency because of unequal Lyonisation. In both G6PD hemizygous males, and homozygous females all enzyme copies are deficient, whereas heterozygous females pattern of the enzymatic deficiency is made up of a mixture of G6PD deficient and normal cells [116]. Like other polymorphisms, this sex-linked gene has been implicated in malaria protection on the basis of geographical correlation suggesting evolutionary selection by malaria [151]. The protection afforded by G6PD deficiency against SM remains debated due to conflicting results. Chapter 3 of this thesis describes the results of my findings from a large case-control study of SM in Kenya.

1.4 Approaches for detecting associations in genetics studies

A wide number of approaches can be used to investigate the relationships between specific genes and malaria risk including twin and adoptee studies, linkage studies, complex segregation analysis studies, candidate-gene studies and genome-wide association studies (GWAS). For the purpose of this thesis, I have only reviewed the last two complimentary approaches (candidate gene and GWAS).

1.4.1 Candidate gene studies

Candidate-gene association studies depend heavily on prior information, usually on particular biological assumptions or the location of the candidate genes within a previously identified region. Polymorphisms found in these genes are tested for statistical association with disease in individuals enrolled in case-control, cohort or family studies. Variants that are found to be susceptible are hypothesized to influence directly an individual's likelihood of developing the disease.

1.4.1.1 Case-control studies

Case-control studies are the most common in the literature. Many authors claim that they are the most powerful and efficient, ensuring robustness when studying a large number of SNPs [152]. A case-control study identifies individuals by

disease status and tests retrospectively for exposure. Specifically, case-control studies in genetic analyses compare the allele frequencies or genotypes in cases (affected individuals) and controls (healthy individuals). Significant differences between the two suggest an association between the alleles being tested and a given disease. Most of the malaria protective genes that have been discovered to date have been investigated using the case-control approach. Case-control studies have the advantage that they are less laborious and expensive than large family studies. For case-control studies, a common method is to use a candidate gene approach but modern techniques and declining costs are increasingly allowing for a more genome-wide and unbiased approach.

1.4.1.2 Cohort studies

Analytically the simplest way to ascertain whether a specific allele or genotype is associated with risk for an infectious disease is by use of a cohort study. Primarily, cohorts of comparable individuals are identified who have and do not have a specific allele or genotype. These individuals are then followed prospectively for a given time to determine whether their exposure affects their risks of getting the disease. The measure of association between exposure and disease is the relative risk. The major drawback of cohort studies is the large sample size leading to high data collection costs.

1.4.2 GWAS approach

Genome-wide association studies seek to identify the vast majority of common loci variants in the human genome that affect disease susceptibility by comparing allele frequencies in individuals with the disease and healthy controls. Since the completion of the Human Genome Project [153, 154] and HapMap [155, 156], genetic epidemiology has entered an era of single nucleotide polymorphisms. With the reduction in the cost and commercial availability of SNP genotyping, many researchers have taken advantage of these technologies, which has led to the publication of hundreds of GWAS seeking to find the genetic causes of many diseases and other traits. For example, the Wellcome Trust Case-Control Consortium [157], conducted association studies of seven common diseases (including bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, Type I and II diabetes) with a total of 14,000 cases and 3000 shared controls whose results were then replicated successfully. Since then, many important GWAS datasets have been generated in African populations, notable among them being studies on malaria conducted by the Malaria Genomic Epidemiology Network (MalariaGEN) [73, 158]. MalariaGEN is a multi-centre initiative that aims to gain a better understanding of the genetic epidemiology of malaria throughout the malaria-endemic world. Besides the likely advantages of using data generated from GWAS, such a

network offers an opportunity for standardizing phenotype definitions, genotyping technology, and an experimental and analytical plan across multiple sites. This is anticipated to augment the power of the study as well as to ensure reproducibility of results, while at the same time allowing the identification and description of regional differences of potential biological interest.

With the above in mind, we are only just beginning to grasp how important host genetic factors are in determining susceptibility to malaria. The majority of previous reported studies have focused on the analysis of single loci; nevertheless, there are instances in which haplotypes rather than genotypes at a single locus can predict the severity of disease [159]. The fundamental difference between haplotypes and individual genotypes at SNPs is that the alleles are inherited together on a chromosome. Haplotypes can be powerful because they yield more information about recombination, which is the physical exchange of DNA molecules during meiosis (cell division). Recombination is important for locating disease-causing mutations by linkage methods and has a great effect on the extent of association between SNPs. Knowledge of the genotype at one SNP can predict the genotype of another SNP if the LD is high between the two SNPs. There has been growing interest in looking at the haplotype structure of a region because it will help in pinpointing the disease-causing locus by marking recombination events in association-based studies. This is particular pertinent in

Africa where the high diversity of the populations results in shorter haplotypes therefore enhancing the likelihood of discovering causative SNPs through LD [160, 161]. In this thesis, I will examine the haplotype structure for two variants and their gene regions that co-exist in Kilifi and further investigate their LD pattern between the neighbouring SNPs.

1.5 Epistasis

In genetics, epistasis is the phenomenon where one gene masks the effect of another [162]. Epistasis occurs if the effect of one locus affecting a complex trait depends on the genotype of a second locus affecting the trait. For instance, consider two loci (A, B), each with two alleles (A, a B, b). Epistasis would occur, for example, if the aabb genotype had a high disease risk, but the eight other possible combinations genotypes had no effect on risk. This is just one of the many possible forms of epistatic interaction between two loci [163]. This definition of epistasis is the same to the concept often used by biologists and molecular geneticist when investigating the interaction between proteins [164].

Epistasis and interaction refer to various aspects of the same phenomenon. The term 'epistasis' is commonly used in population genetics and refers specifically to the statistical properties of the phenomenon, and does not automatically imply a biochemical interaction between gene products. However, in broad-spectrum, epistasis is used to signify the departure from 'independence' of the effects of different genetic loci. Confusion often arises due to the heterogeneous interpretation of 'independence' between various branches of biology. For further discussion of the definitions of epistasis, and the history of these definitions, see [164]. For the purposes of this thesis the term epistasis remains more general;

that is, the risk of having a phenotype can increase or decrease as a result of combination of two or more alleles/genes.

The presence of epistasis can have key implications for the interpretation of statistical models. If two loci of interest interact, the relationship between each of the interacting loci and a phenotype of interest depends on the genotype values of the two loci. In practice, this makes it more harder to predict the consequences of changing the value of a variable, especially if the variables it interacts with are hard to measure or difficult to control [165]. It is increasingly clear, therefore, that we will need to consider epistasis when we explore genotype-phenotype relationships. Failure to recognize these interactions could obscure important genetic loci.

A substantial number of studies in complex human diseases, including diabetes [166, 167] and some forms of cancer [168, 169], have shown susceptibility genes that contribute to a common disease trait which interact significantly with one another in combined analyses [170, 171]. Several investigators have found alleles that have opposite effects depending on the genetic background [172, 173], which further raises the likelihood of overlooking epistatic susceptibility genes in single-gene analyses [174]. Other concrete examples of epistasis between genes exist such as interactions between Apolipoprotein E (*ApoE4*) and the Low-density

Lipoprotein Receptor (*LDLR*) genes in coronary artery disease [175] and identification of gene-gene interactions for asthma susceptibility loci in three U.S. populations [176].

1.5.1 Epistasis between malaria protective genes

To date, only a small number of epistatic interactions have been investigated for malaria-protective polymorphisms. Most researchers have focused mainly on the effect of individual polymorphisms that contribute to the disease phenotype. However, it has been shown that co-inheritance of several polymorphisms and interactions between them may also play significant roles in determining phenotype.

1.5.1.1 HbAS and α^+thalassaemia

In a study conducted in Kilifi, Williams and colleagues [103] found evidence for epistasis in the malaria-protective effect of two important haemoglobin variants: HbAS and α^+thalassaemia [103]. Both variants were independently associated with a protective effect against uncomplicated malaria as detected through an active community surveillance and admission to hospital with SM [103]. Nonetheless, this protective effect afforded by the HbAS and α^+thalassaemia when co-inherited together was lost (see Figure 1.6) [103]. This negative epistasis

has subsequently been confirmed in two separate populations from Ghana and Mali [57, 62].

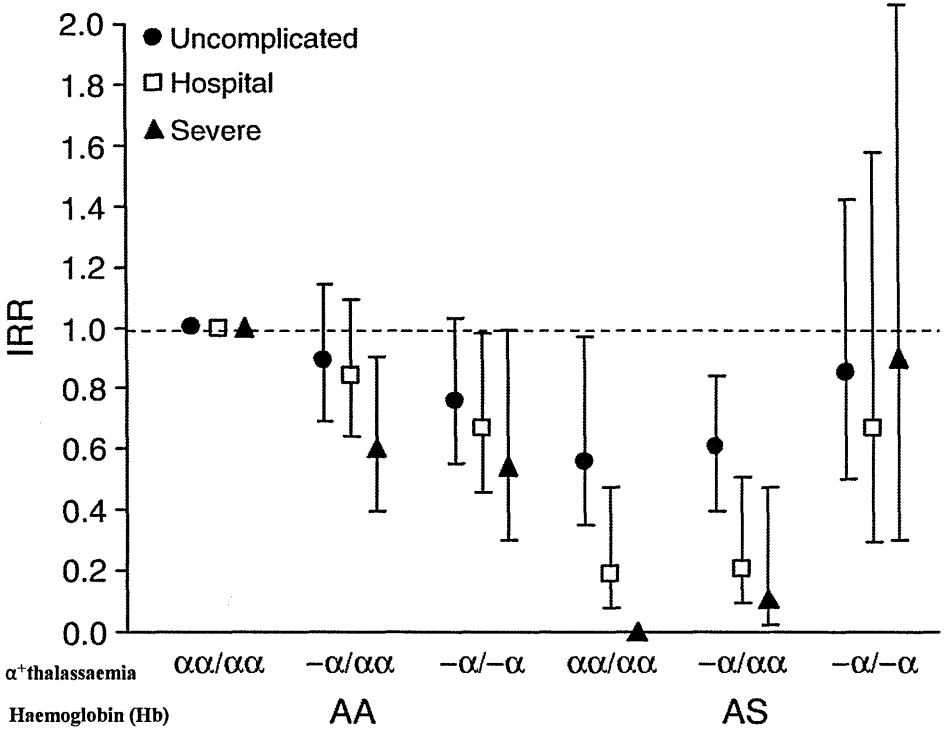
1.5.1.2 Haptoglobin and α^+thalassaemia

In a more recent study conducted in Kilifi, Atkinson and colleagues found evidence for epistasis between haptoglobin (Hp) genotype and α^+thalassaemia variant[177]. The protective effects against SM afforded by α^+thalassaemia variant alone and for α^+thalassaemia when co-inherited by Hp type are summarized in Figure 1.7. The study observed increased risk associated with Hp2-2 genotype group in the presence of α^+thalassaemia (Figure 1.7) [177].

1.5.1.3 HbAS and G6PD deficiency

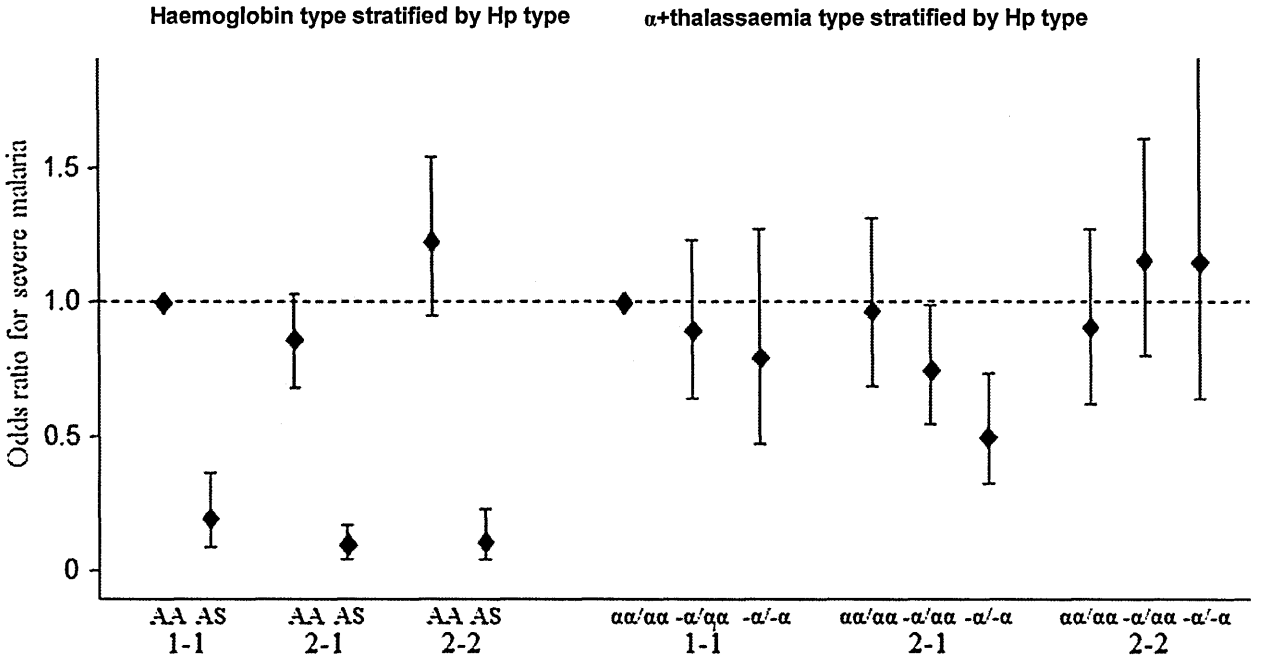
There is also some evidence of a negative interaction between HbAS and G6PD (A-) deficiency from a case-control study in Mali [178] where both HbAS and G6PD (A-) deficiency are common, occurring at frequencies of 12% and 14% respectively in the population. Overall, the study observed a protective effect for HbAS among females against SM (OR, 0.33; P=0.01); however, rather than experiencing additive protection, co-inheritance of HbAS and heterozygous G6PD (A-) deficiency in females was associated with an increased risk against SM (OR, 15.00; P=0.003). However, given the small number of subjects with these two polymorphisms the study results should be treated with caution.

Figure 1.6. Incidence rate ratios for malaria by haemoglobin type and α^+thalassaemia genotype showing a negative epistatic interaction between the two variants.



Compared with the baseline group (children with both HbAA and normal thalassaemia genotype,αα/αα), the incidence was lowest in HbAS children with αα/αα; no additional advantage derived from co-inherited −α/αα, and the protection afforded by HbAS was lost with co-inherited −α/-α. Adapted from Williams *et al.* 2005 [103], and published with permission from Nature Publishing Group.

Figure 1.7. The Odds ratio for severe malaria by haemoglobin type and α⁺thalassaemia type stratified by haptoglobin genotypes.



Logistic regression models were used to determine the ORs for risk of SM by Haemoglobin type (AA, AS) and α⁺thalassaemia type (αα/αα, −α/αα, −α/−α) stratified by haptoglobin genotypes (1-1, 2-1, 2-2). The AA, 1-1 and αα/αα were the respective reference types with an OR for SM of 1.0. The risk of SM was lowest for −α/−α thalassaemia inherited in combination with 2-1, but α⁺thalassaemia was not protective when inherited in combination with 2-2. The 1-1 genotype was associated with a non-significant trend toward protection against SM when inherited with α⁺thalassaemia. Adapted from Atkinson *et al.* 2014 [177], and published with permission from the American Society of Haematology.

1.5.2 The challenge of detecting epistasis

Unlike a number of microorganisms such as budding yeast, for which epistatic interactions have been carefully studied through synthetic models [179], humans are much more challenging subjects to study. This is especially due to the high number of genes involved and the impracticability of conducting direct assays on humans (for instance, we cannot directly modify an individual's DNA). As a result, analysing natural variation in genomic measurements has been the most feasible method for finding epistasis in humans. Nevertheless, this approach also poses a number of challenges, as stated below.

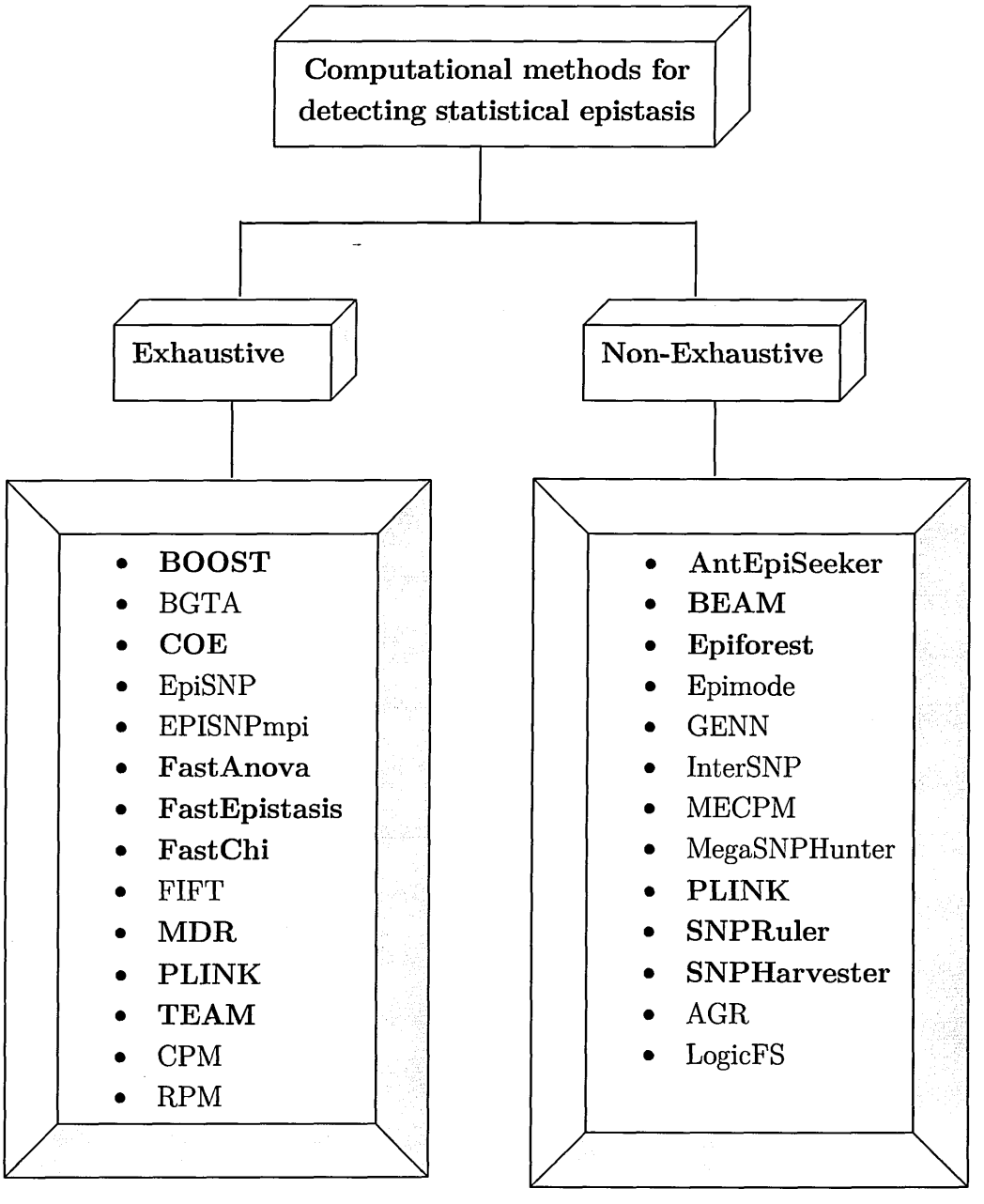
In order to detect epistasis in a given analytical dataset, one needs to find a pair or a set of markers that shows statistically significant epistatic interaction. Nevertheless, the number of combinations to consider increases very quickly: enumerating all possible pair-wise interactions among N markers requires testing N (N-1)/2 combinations. In the case of a GWAS study of 500,000 markers (for example), this requires 124,999,750,000 tests, this being just the tip of the iceberg if three-way interactions were to be considered. Therefore, even with the existence of large computing clusters, exhaustively searching for epistasis will be computationally challenging. In addition, testing a large number of hypotheses leads to a reduction in the statistical power of detecting an epistatic effect.

Multiple testing correction refers to a procedure that proposes stricter significance thresholds to counteract the rise of false positives caused by testing a large number of hypotheses. However, in detecting epistasis, the number of hypotheses is so large that any attempt to control the number of false positives comes at an expense of possibly losing many epistatic interactions.

1.5.3 Novel approaches for detecting epistasis

Cordell [164] provides a recent and thorough review of current epistasis analysis methods published in the literature as well as the potential benefits and drawbacks of each. The methods differ according to whether one is carrying out an association or linkage analysis and according to whether one is dealing with a 'quantitative' or a 'qualitative' trait. For genetic association studies, standard methods for epidemiological studies may be employed, with genotypes at various loci considered as risk for a disease. As per the literature, commonly used algorithms for detecting epistasis in human complex disease are divided into two broad categories: those that explicitly test every possible interaction up to some size **(exhaustive)** and those that avoid an exhaustive enumeration of the search space **(non-exhaustive).** Figure 1.8 shows a diagram summarising some of the methods that can be found in the literature.

Figure 1.8. Classification of the methods that detect statistical epistasis.



Most methods can be classified into two categories according to their search strategies: exhaustive search and non-exhaustive search. Methods in bold are described in detail in **Appendix A, Table A.1** while methods in blue are tested and evaluated in detail in Chapter 4.

1.5.3.1 Exhaustive Algorithms

An "exhaustive" method is one that enumerates all possible k-way interactions for some k in order to identify the effect or effects which best predict phenotypic outcomes. Some methods go even further, testing every possible partitioning of alleles. As the size of available data sets exceeds one million markers, it is evident that this search for interactions will present computational and statistical difficulties and that feasibility and the need to control for multiple testing needs to be taken into account.

1.5.3.2 Non-Exhaustive Algorithms

If an exhaustive method is one that searches all possible k-way interactions, a "non-exhaustive" algorithm performs a partial search of the possible interaction space to terminate relatively quickly. While they are typically faster than exhaustive procedures, it is impossible to know if any such method will identify or even test the correct solution for any given data set. Non-exhaustive algorithms can be further classified according to their search space reduction strategy. **Greedy** methods perform filtering based on non-epistatic or lower-order interaction results to filter markers displaying no main or low-order effects. The success of the greedy strategy depends on the nature of interactions present in the data set: pure epistatic interactions displaying no main effects are likely to

be missed. **Stochastic** algorithms iteratively select a small number of loci and perform a thorough test for epistasis. This strategy relies on luck to select interacting loci in at least one iteration.

In summary, algorithms to detect multi-locus genotype-phenotype association may be classified as exhaustive or non-exhaustive. Members of the former category test all possible interactions up to a user-specified size, while the others use a greedy heuristic or stochastic search strategy to quickly identify causative loci. While it is beyond the scope of this thesis to describe and test all published methods, I have reviewed some common methods available in the literature and enumerated their advantages and disadvantages in detecting SNP–SNP interactions in **Appendix A, Table A.1**. I have selected two representative algorithms: PLINK and AntEpiSeeker to explore in depth using the candidate gene case-control study. These, along with a novel method, are evaluated for statistical performance and computational efficiency in Chapter 4.

1.6 Motivation for this work

Despite various attempts to control malaria, the disease killed more than half a million people in 2013, the majority being children under the age of five years old in SSA [2]. The intriguing phenomenon about malaria is that although the majority of children who are exposed to repeated infection survive, only a

relatively small proportion develops severe and life-threatening complications of the disease. Host genetic factors are thought to be important modifiers of severe and fatal malaria [34], yet insufficient is known about which genes are involved or the type or degree of protection that they might afford.

A wealth of host genetic association studies on malaria exists whose results have not been replicated in other settings. Given the complex nature of malaria and a number of confounding factors these variations in results are expected. One of the main reasons for this lack of reproducibility is the lack of statistical power in genetic association studies to detect these associations.

Current trends in the field of genetics are, however, revolutionizing genetic studies. A number of studies have revealed the complex genetics of malaria susceptibility and a number of genes have already been associated with malaria susceptibility even though the sample-size of these studies has often been too small. The HbAS and α^{+}thalassaemia are the best studied of these, showing the greatest protection [180] and numerous mechanisms of protection having been proposed [181]. However, the reality that these polymorphisms are commonly co-inherited, has received little attention [103]. Apart from the interactions seen between HbAS and α^{+}thalassaemia, are there other important polymorphisms

that interact with HbAS or α^{+}thalassaemia to influence malaria outcome? Do pure interactions really exist in genetics?

With the sequencing of the human genome [153] and high-throughput genotyping technologies now available, researchers now have a set of tools that make it easier to find the genetic contributions to complex diseases such as malaria. As shown by Mackinnon *et al* [33, 34], the genetic basis of resistance to malaria is expected to involve many different malaria candidate genes, individually resulting in only small population effects, which might be missed due to methodological issues such as sample size.

In recent years, I have personally gained considerable experience in my current role as a MalariaGEN Data Fellow including how to prepare and analyse genetic data and how to conduct a number of different bioinformatics and statistical analyses. I have spent some of my efforts in developing a resource for analysing the associations between published genetic markers and SM in children in the study area. This data will not only be the first and most detailed genetic data for Kilifi, but also serve as quality control data for further studies to be conducted in the area. In addition the study will shed more light on the replication or non-replication of associations, and may also come up with new functional SNPs or markers in LD with functional SNPs for the study area.

Another great advantage of collaborating with MalariaGEN is that the network has access to a large archive of DNA samples from a large number of malaria-endemic countries that have ethical approval for genetic studies and this will allow me to use a large dataset such as GWAS.

To summarise the discussion above, precedence for the search for host genetic factors thought to be important modifiers of SM exists, yet not enough is known about which genes are involved or the type or degree of protection that they might afford. This thesis seeks to address this question from four main aspects: reappraise a range of candidate malaria-protective genes using a large-scale case-control study of SM; compare different approaches of detecting epistasis interactions and also look for evidence of these interactions between candidate genes; examining the haplotype structure and LD for two important variants that co-exist in Kilifi and finally to use them in a conditional genome-wide association interaction study as a starting point to investigate the process of detecting epistasis in SM. To summarise the objectives of the proceeding chapters:

1. Chapter 2 discusses material and methods applied throughout the thesis.
2. Chapter 3 reappraise a range of candidate malaria-protective genes using a large-scale case-control study of SM.

3. Chapter 4 compares different approaches of detecting epistasis interactions, and also examines evidence of these interactions between candidate genes.
4. Chapter 5 explores the complex haplotype structure and LD patterns for two important variants which are found to co-exist in Kilifi.
5. Chapter 6 elucidates a methodological framework for detecting epistatic interaction using a genome-wide case-control study
6. Finally, in Chapter 7 general conclusions are drawn and future work that may follow this thesis work are summarised.

In the appendices, some additional materials including pseudo-codes that might be of interest to the reader are presented.

Chapter 2

Materials and Methods

2.1 The study area and study site

The study was conducted in Kilifi County on the Indian Ocean coast of Kenya. The economy of the area is predominantly rural, being based on subsistence farming of maize, cassava, cashew nuts and coconuts as well as dairy cows and goats. In addition, two large agricultural estates, two research institutes and several tourist hotels contribute to local employment. The majority of the population falls within the Chonyi, Giriama and Kauma sub-tribes of the Mijikenda ethno-linguistic group.

The County is traditionally a malaria endemic area, transmission being seasonal following long and short rainy seasons from April-June and October-November respectively. The vast majority of malaria infections in the area are caused by *P. falciparum*. Malaria is predominantly transmitted by mosquitoes of the species *A. gambiae* but *A. funestus* may also play a minor role [182, 183] with an overall EIR estimated at 1-100 infective bites / year [182, 184]. Over the last two

decades, significant declines have been observed both in malaria transmission and in the incidence of SM [185, 186]. The present study was conducted at the KEMRI / Wellcome Trust Research Programme (KWTRP) in Kilifi which is based in the grounds of the Kilifi County Hospital (KCH), the first referral centre for the majority of the population of Kilifi County. The KWTRP was established in 1989 and its research interests centre around important child health issues [187].

2.1.1 The Kilifi health and demographic surveillance system (KHDSS)

The KHDSS was established in September 2000, with the primary aim of providing an enabling infrastructure for the conduct of multiple, large scale, interdisciplinary epidemiological studies. Many but not all of our studies involve an element of case-ascertainment at the KCH. For this reason, the geographic limits of the KHDSS are based on the pattern of admissions from the community surrounding KCH. Approximately 80% of patients admitted to KCH reside in the study area (Figure 2.1 right panel). The KCH admits between 4000 and 5000 paediatric patients and 3-4000 adult patients each year [187]. The KHDSS is a member of the International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH) [188], which is a network of demographic research centres within the developing world [188].

Figure 2.1. A Map of Kilifi County on the Coast of Kenya showing the Kilifi Health and Demographic Surveillance System and admission rates to Kilifi County Hospital.

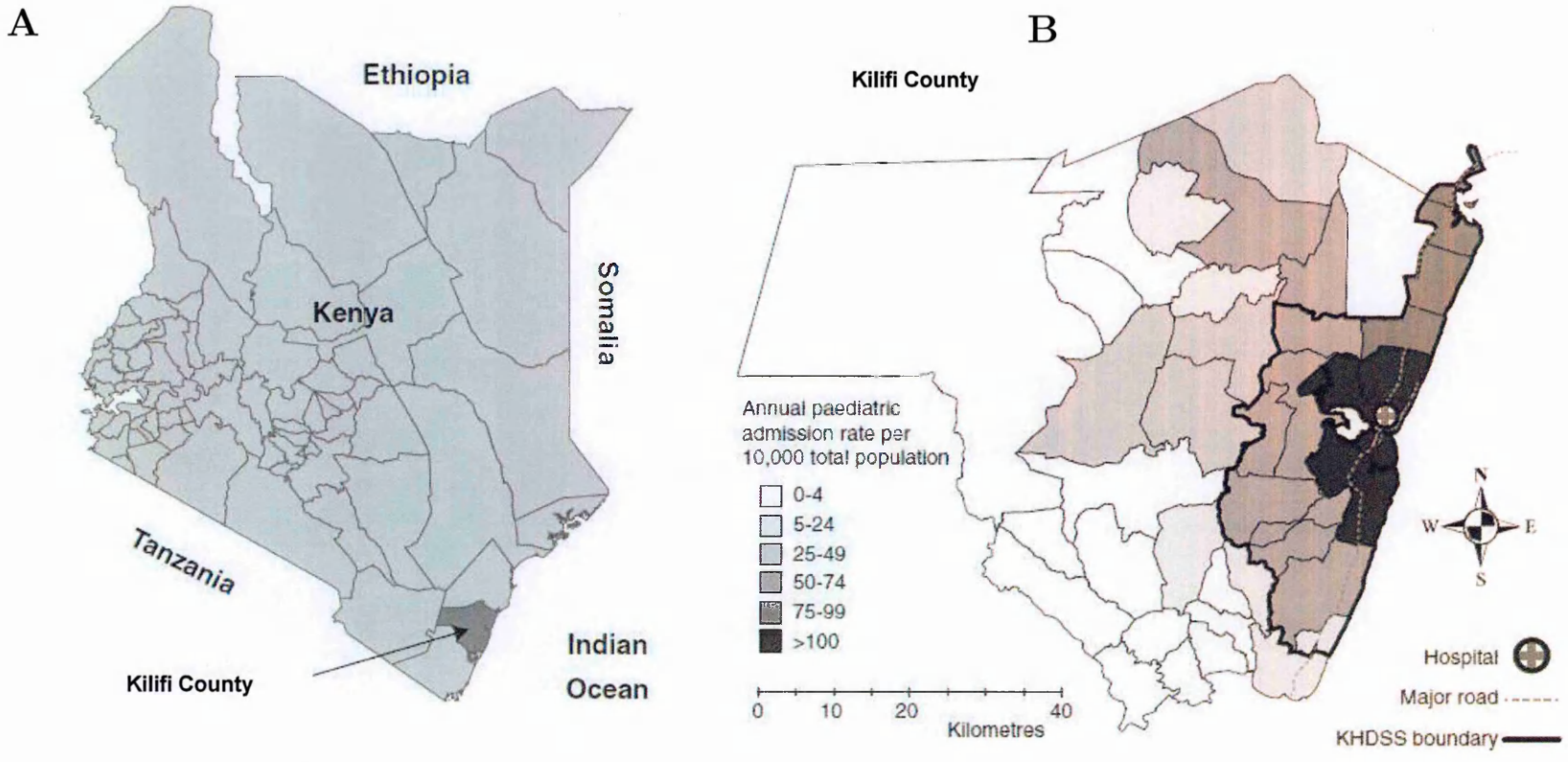


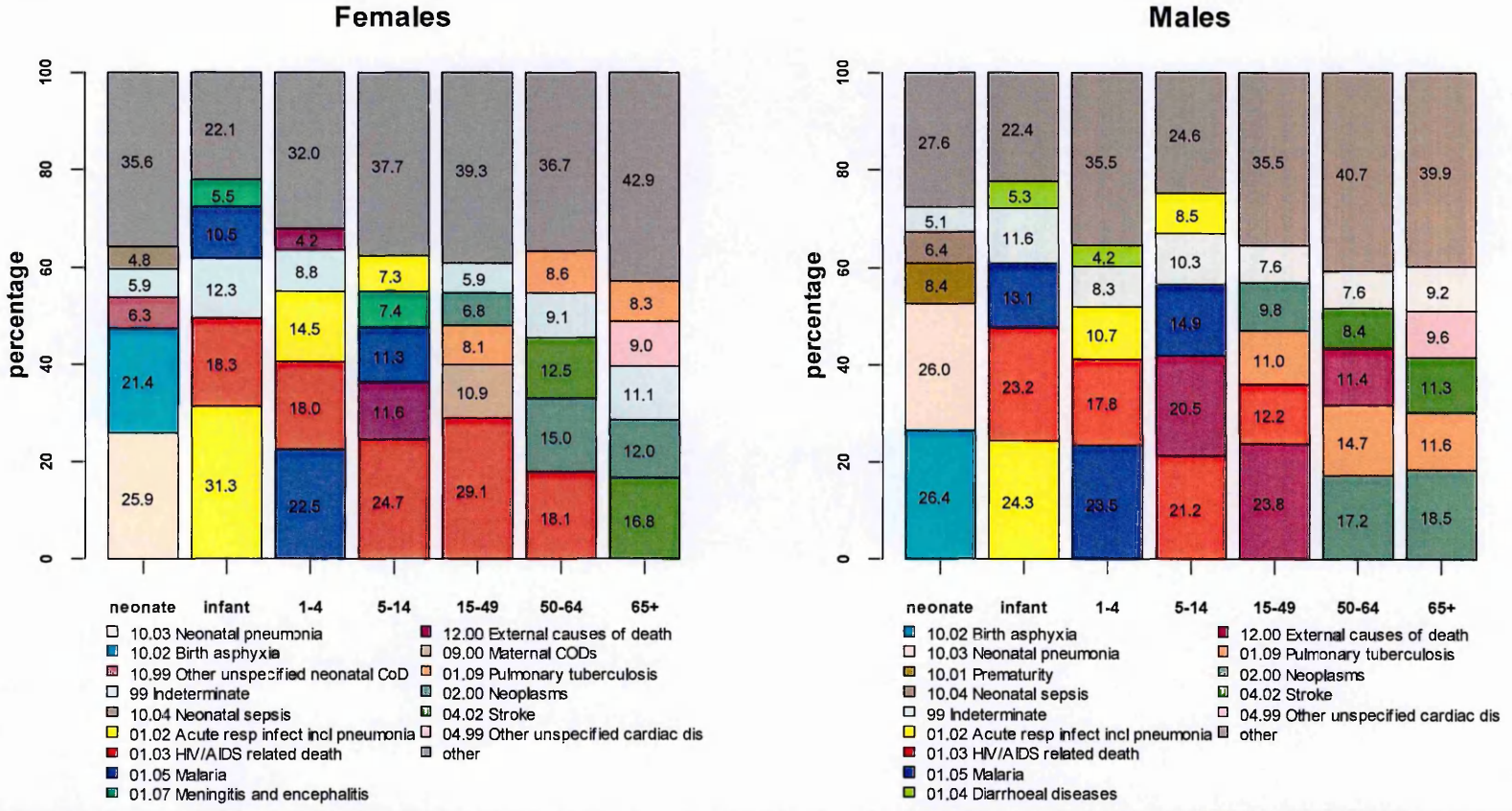
Figure 2.1 a) A map of Kenya showing neighbouring countries and location of the Kilifi County (indicated by the black arrow). Figure 2.1 b) The map of Kilifi County showing the rate of paediatric admissions to Kilifi County Hospital (marked by the Hospital sign) stratified by administrative sub-locations. Adapted from Scott *et.al* 2012 [187].

The KHDSS covers an area of 891km^2 which is divided into 15 administrative locations, 40 sub-locations and 186 enumeration zones. The current population of the KHDSS area is approximately 280,000 [187]. Every three months the data are updated with births, deaths, pregnancies and migration events through household visits by enumerators. A critical feature of the KHDSS, that facilitates many of the KWTRP objectives, is that all children who present to the wards and clinics at KCH are identified in real-time on a central computer-based database using a unique personal identification number (PID). This feature facilitates research conducted by specific studies within the KWTRP since individuals may be traced through all databases and basic demographic characteristics such as sex, age, and location of residence can easily be retrieved including results of a range of routine laboratory investigations. For instance linkages between the KHDSS and the paediatric ward have allowed the KWTRP researchers to evaluate the impact of insecticide-treated bed-nets on malaria morbidity and mortality in the community surrounding the study area [189].

The KHDSS registers 1200 to 1500 deaths within the resident population every year. More than 60% of these deaths occur outside the hospital where the causes of death are rarely recorded. In collaboration with the Kenyan Ministry of Health (MOH) [190] in 2008, the KHDSS began collecting verbal post-mortems of all the deaths reported by field workers with the aim of documenting the

pattern of underlying cause of death (COD) at community level using a verbal autopsy (VA) tool [191]. The pattern of deaths within the KHDSS, attributed according to VA diagnoses, is summarized in Figure 2.2. Of particular interest, malaria was the main COD in children aged 1-4 years, despite the reported decline in malaria admissions in previous studies conducted in Kilifi [192, 193].

Figure 2.2. Distribution of the top five causes of death among the residents of Kilifi Health Demographic Surveillance System stratified by age group and gender, N=4460.



Causes of death (COD) were derived using a verbal autopsy approach which is a method of finding out the cause of a death based on an interview with next of kin or other caregivers. The x-axis shows different age groups while y-axis shows the percentages. A different colour depicts a different COD. Substantially high percentage of malaria deaths (the bars in blue) were contributed by young children aged 1-4 years. Adapted from Ndila *et.al* 2014 [194].

2.2. The study design

In the present study I have utilised two major classes of genetic association study designs – candidate gene and genome-wide – with an emphasis on how each has contributed to our understanding of the genetics of SM in Kenya. Both studies were conducted as collaboration between the KWTRP and the MalariaGEN consortium.

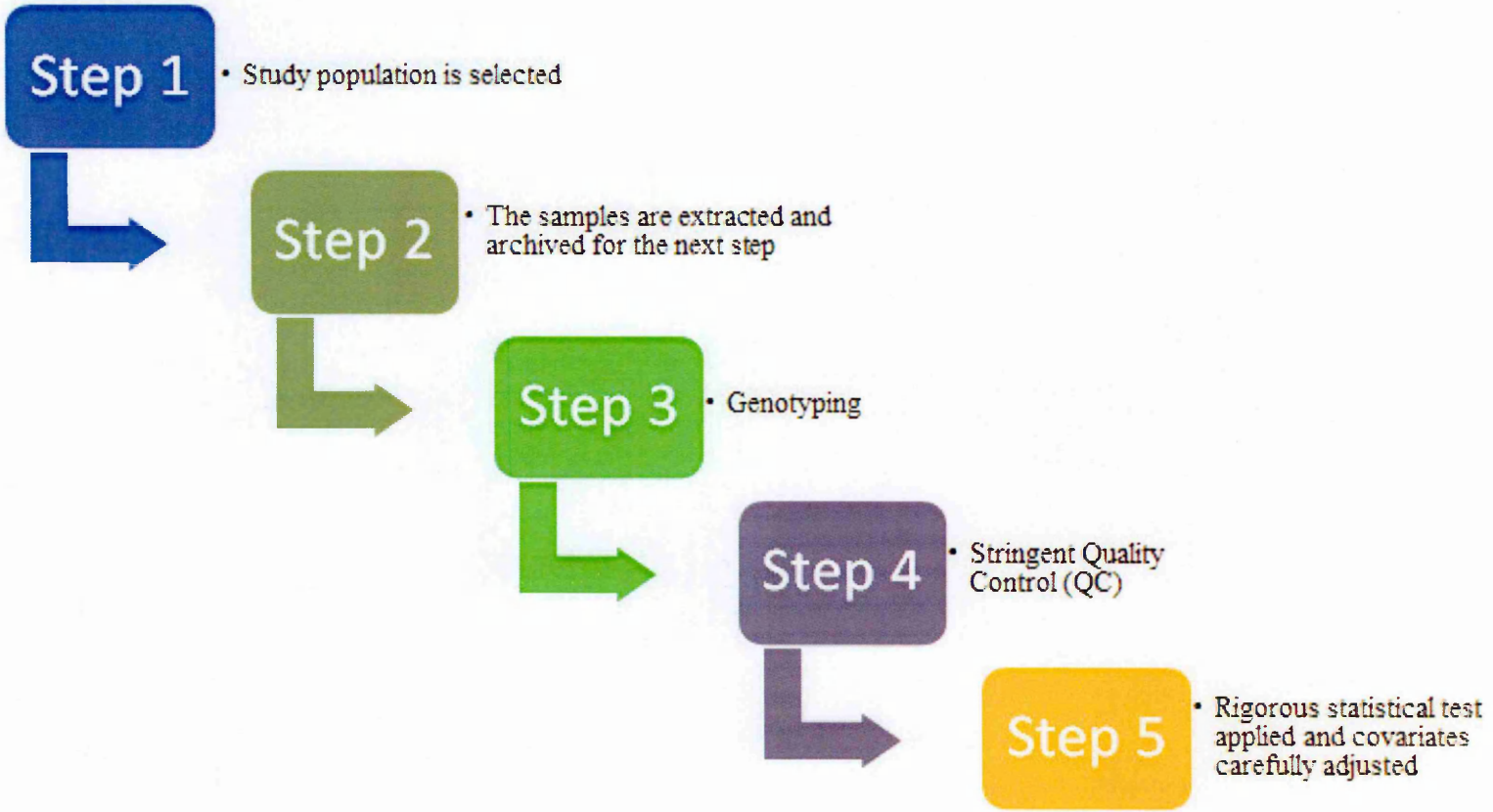
2.2.1 A candidate gene case-control study

As the first step to elucidate the pathology of complex diseases such as malaria, a study of candidate genes provides various clues to understand genes responsible for the phenotype of interest. As shown in Figure 2.3, the general workflow of the current study was divided into five steps:

1. Study population
2. DNA extraction and archiving
3. Genotyping
4. Quality control
5. Statistical test analysis

They are all described in the subsequent sections.

Figure 2.3. A typical workflow of the candidate gene case-control study.



2.2.1.1 Study population

SM cases were recruited as part of on-going surveillance studies of SM at KCH. Cases were children aged 0-13 years who were admitted to the high dependency unit (HDU) at KCH between June 1995 and June 2008 with *P. falciparum* parasitaemia and one or more clinical features of SM, as described in detail in Chapter 1 (section 1.1.3.1). All cases were treated according to standard guidelines [195]. For the definition of SM sub-phenotypes, in addition to the presence of P. falciparum malaria parasites in the peripheral blood, detected by microscopy, I defined CM as a Blantyre Coma Score of ≤2, SMA as a haemoglobin concentration of <5g/dl or a packed cell volume of <15%, and RD as the presence of abnormally deep breathing. A number of children had combinations of these phenotypes. A subset of children admitted with severe *P. falciparum* infection did not present with CM, SMA or RD but suffered from hypoglycaemia, prostration or metabolic acidosis. They were therefore classified as a separate category and described as "Other SM". For the purpose of this thesis only SMA, CM and RD sub–phenotypes were analysed.

Controls were children 3-12 months of age who were born between August 2006 and September 2010 within the same study area as cases, and who were recruits to an ongoing cohort study investigating genetic susceptibility to a range of childhood diseases [196].

2.2.1.2 DNA extraction and archiving

Blood samples were collected on admission to KCH for the SM cases while for the healthy controls, blood samples were collected during home visits by the fieldworkers. Extraction of genomic DNA was done using standard methods; either from fresh whole blood samples or from blood samples previously stored at -80°C using proprietary methods ABI PRISM (Applied Biosystems, California, USA) or Qiagen DNA Blood Mini Kit™ (Qiagen, West Sussex, United Kingdom). Aliquots of the extracted DNA samples were sent to the MalariaGEN Resource Centre in Professor Kwiatkowski's Laboratory at the Wellcome Trust Centre for Human Genetics (WTCHG), Oxford, UK and archived. The archiving process entailed re-labelling of samples with new sample identity numbers (ID's) for confidentiality, determining the concentration of samples using the PicoGreen® double strand (dsDNA) DNA Quantification Kit (Molecular Probes, Inc) and then placing samples in boxes before storage at -80°C. Finally, all the Meta-data regarding the samples was stored into a secure web-based laboratory information management system (LIMS) known as 'Topheno' established by the MalariaGEN. Topheno captured essential clinical information for individuals in the study (but not personal data) and allowed for the conversion of variables to standard units (for example gender coded as male or female and parasitaemias converted to parasites per microlitre).

2.2.1.3 Genotyping

In order to increase the amount of DNA required for high-throughput genotyping, all DNA samples underwent whole genome amplification through either Primer Extension Pre-amplification (PEP) Polymerase Chain reaction (PCR) using 15N base primers [197] or Multiple Displacement Amplification (MDA) [198], before genotyping on the Sequenom® MassARRAY iPLEX™ platform (Agena Biosciences, Hamburg, Germany) [117, 199]. More details on the genotyping protocol have been described in detail in [45].

In the current study a total of 136 SNPs in 71 malaria candidate genes including α^{+}thalassaemia were genotyped. These SNPs were chosen based on previously published reports of malaria candidate-gene associations in addition to SNPs that had potential associations in GWAS studies undertaken by the MalariaGEN consortium [73, 200]. Additional assays were designed for gender determination by comparing the Amelogenin gene between the X and Y chromosomes (see attributions). DNA samples were also typed for α^{+}thalassaemia (the common African 3.7-kb α-globin deletion) in Kilifi and at WTCHG (see attributions) by a modified multiplex PCR method as described previously [41]. See Table 2.1 for details of the genes/SNPs included in this study.

Table 2.1. The candidate genes selected for genotyping in the case-control study.

#	Gene symbol	Gene description	SNP name	Chr	Position
1	*AJAP1*	Adherens Junctions Associated Protein 1	kgp15825649	1	4817664
			rs6674631	1	4834821
2	*LPHN2*	Latrophilin 2	rs146428334	1	81688714
			rs72933304	1	81726138
			rs72933310	1	81727427
			rs72933350	1	81751439
			rs4650365	1	81770827
3	*GBP7*	Guanylate Binding Protein 7	rs1803632	1	89582690
4	*DARC*	Duffy Antigen Receptor for Chemokines	rs2814778	1	159174683
5	*ATP2B4*	Atpase, Ca (2+) Transporting, Plasma Membrane 4	rs55868763	1	203652140
			rs1541255	1	203652141
			rs10900585	1	203654024
			rs4951074	1	203660781
			rs3753036	1	203677250
6	*IL10*	Interleukin 10	rs3024500	1	206940831
			rs1800896	1	206946897
			rs1800890	1	206949365
7	*CR1*	Complement Receptor 1	rs17047660	1	207782856
			rs17047661	1	207782889
8	*LOC727982*	-	rs1371478	2	4901589
			rs1371474	2	4909777
			rs10188961	2	4926593
9	*LAPTM4A*	Lysosomal Protein Transmembrane 4 Alpha	rs973128	2	20332487
10	*SDC1*	Syndecan 1	rs11899121	2	20367973
11	*IL1A*	Interleukin 1, Alpha	rs17561	2	113537223
12	*IL1B*	Interleukin 1, Beta	rs1143634	2	113590390
13	*ZSWIM2*	Zinc Finger, SWIM-type containing 2	rs4316902	2	188007364
			rs144778284	2	188012821
14	*IL17RE*	Interleukin 17 Receptor E	rs708567	3	9960070
15	*OXNAD1*	Oxidoreductase NAD-Binding Domain Containing 1	kgp9483807	3	16407519
			rs79691057	3	16408251
			rs75180423	3	16408723
16	*TLR9*	Toll-Like Receptor 9	rs187084	3	52261031
17	*IL17RD*	Interleukin 17 Receptor D	rs6780995	3	57138419
18	*ARL14*	ADP-Ribosylation Factor-Like 14	rs76033371	3	160362359
			rs75731597	3	160364808
			rs74954675	3	160381509
19	*B3GALNT1*	Beta-1,3-N-Acetylgalactosaminyltransferase 1 (Globoside Blood Group)	rs12107243	3	160793678
20	*TLR1*	Toll-Like Receptor 1	rs4833095	4	38799710
21	*TLR6*	Toll-Like Receptor 6	rs5743810	4	38830350
			rs5743809	4	38830514
22	*INPP4B*	Inositol Polyphosphate-4-Phosphatase, Type II	rs77389579	4	143538511
			rs13103597	4	143558581
23	*USP38*	Ubiquitin Specific Peptidase 38	rs4266246	4	143971242

			rs28459062	4	144039139
24	*GAB1*	GRB2-Associated Binding Protein 1	rs7663712	4	144261117
25	*GUSBP5*	Glucuronidase, Beta Pseudogene 5	rs148111931	4	144540045
26	*FREM3*	FRAS1 Related Extracellular Matrix 3	rs184908374	4	144665753
			rs149914432	4	144666678
			rs186790584	4	144680140
			rs184895969	4	144698528
			rs186873296	4	144702474
27	*GYPB*	Glycophorin B (MNS Blood Group)	rs191338817	4	144948956
28	*C6*	Complement Component 6	rs1801033	5	41199959
29	*IRF1*	Interferon Regulatory Factor 1	rs2706384	5	131826880
30	*IL13*	Interleukin 13	rs20541	5	131995964
31	*IL4*	Interleukin 4	rs2243250	5	132009154
32	*GABBR1*	Amma-Aminobutyric Acid (GABA) B Receptor, 1	rs192151845	6	29588309
33	*HCG4*	HLA Complex Group 4	rs114980857	6	29772098
34	*LTA*	Lymphotoxin Alpha	rs2239704	6	31540141
			rs909253	6	31540313
35	*TNF*	Tumour Necrosis Factor	rs1799964	6	31542308
			rs1800629	6	31543031
			rs361525	6	31543101
			rs3093662	6	31544189
36	*HSPA1B*	Heat Shock 70kda Protein 1B	rs6457452	6	31795550
37	*SNORD48*	Small Nucleolar RNA, C/D Box 48	rs116288147	6	31803074
38	*CTL4*	-	rs2242665	6	31839309
39	*IL20RA*	Interleukin 20 Receptor, Alpha	rs1555498	6	137325847
40	*PLEKHG1*	Pleckstrin Homology Domain Containing, Family G Member 1	rs55958968	6	150942218
			rs144224092	6	150973623
			rs79100774	6	150975934
			rs114726617	6	150980481
			rs2131263	6	150981102
			rs76924464	6	150982529
			rs151293197	6	150994429
			rs142712208	6	151026346
			rs15116938	6	151046029
			rs14155519	6	151048708
41	*NOD1*	Nucleotide-binding Oligomerization Domain containing 1	rs2075820	7	30492237
42	*CD36*	Cluster Of Differentiation 36	rs3211938	7	80300449
43	*TLR4*	Toll-Like Receptor 4	rs4986790	9	120475302
			rs4986791	9	120475602
44	*ABO*	ABO Blood Group	rs150311214	9	136131057
			rs56390333	9	136131064
			rs8176746	9	136131322
			rs8176719	9	136132909
45	*MKI67*	Marker of proliferation Ki-67	rs11016116	10	129975450
			rs148494166	10	129976030
			rs115947774	10	130072795
46	*RHOG*	Ras Homolog family member G	rs138826089	11	3847190
47	*RRM1*	Ribonucleotide Reductase M1	kgp12768002	11	4111415
48	*HBB*	Hemoglobin Beta Chain	rs334	11	5248232

49	*TRIM5*	Tripartite Motif Containing 5	rs7935564	11	5718517
50	*RTN3*	Reticulon 3	rs542998	11	63487386
51	*GRIP1*	Glutamate Receptor Interacting Protein 1	rs192909543	12	67366471
			rs1394263	12	67366537
52	*CAND1*	Cullin-Associated and Neddylation-Dissociated 1	rs1566830	12	67369898
			rs12307123	12	67394950
			rs10459266	12	67455888
53	*IL22*	Interleukin 22	rs2227507	12	68642647
			rs1012356	12	68644618
			rs2227491	12	68646521
			rs2227485	12	68647713
			rs2227478	12	68648622
54	*TPTE2*	Transmembrane Phosphoinositide 3-Phosphatase and Tensin Homolog 2	rs182873742	13	20050239
55	*SPTB*	Spectrin, Beta, Erythrocytic	rs229587	14	65263300
56	*LTBP2*	Latent Transforming Growth Factor Beta Binding Protein 2	rs74063230	14	75066093
57	*YLPM1*	YLP Motif Containing 1	rs10139016	14	75274288
58	*RPS6KL1*	Ribosomal Protein S6 Kinase-Like 1	rs3742785	14	75373034
59	*ADCY9*	Adenylate Cyclase 9	rs2230739	16	4033436
			rs10775349	16	4079823
60	*HBA*	Hemoglobin, Alpha	α^{+}thalassaemia	16	222846
61	*IL4R*	Interleukin 4 Receptor	rs1805015	16	27374180
62	*ADORA2B*	Adenosine A2b Receptor	rs2535611	17	15861332
63	*NOS2*	Nitric Oxide Synthase 2A (Inducible, Hepatocytes)	rs2297518	17	26096597
			rs1800482	17	26128509
			rs9282799	17	26128728
			rs8078340	17	26129212
64	*BCAS3*	Breast Carcinoma Amplified Sequence 3	rs184142841	17	58855323
65	*TBX2*	T-Box 2	rs73991577	17	59323072
66	*EMR1*	Egf-Like Module Containing, Mucin-Like, Hormone Receptor-Like 1	rs373533	19	6919624
			rs461645	19	6919753
67	*ICAM*	Intercellular Adhesion Molecule 1	rs5498	19	10395683
68	*GNAS*	Guanine Nucleotide Binding Protein, Alpha	rs8386	20	57485812
69	*DERL3*	Derlin 3	rs1128127	22	24179132
70	*CD40LG*	CD40 Ligand (TNF Superfamily, Member 5)	rs3092945	X	135729609
			rs1126535	X	135730555
71	*G6PD*	Glucose-6-Phosphate Dehydrogenase	rs1050829	X	153763492
			rs1050828	X	153764217

All SNPs are referenced to GRCh37, dbSNP135 and Ensemble Build 68. Abbreviations: SNP=Single nucleotide polymorphism; Chr=Chromosome.

2.3.2.3 Quality control process

The genotype data obtained from the assays are raw data, containing various errors due to sample quality or procedural issues. These errors may lead to spurious findings; thus a set of quality control (QC) procedures are required. A stringent QC process was used to inform the selection of appropriate SNP and sample inclusion criteria.

1. Obtaining sample pass rate (the proportion of samples for which the genotypes of the SNP were successfully determined) and assay pass rate (the proportion of SNPs that were successfully genotyped) for each SNP. The two were assessed using the empirical cumulative distribution function and an *ad hoc* cut-off that sensibly removed outlying SNPs or samples was deployed. All samples or SNPs with missingness (the presence of undefined genotypes) >10% were excluded.
2. An allele frequency threshold of 5% was applied to each SNP. All SNPs with frequencies of <5% were removed because of a low statistical power.
3. One method to assess genotyping quality included testing the conformation of the observed genotype distributions in the controls to the expected distributions under Hardy-Weinberg equilibrium (HWE) using a *Chi*-square (χ^2) statistical test, with one degree of freedom. The HWE principle states that in a population large enough and in the absence of selection, migration,

and mutation, the frequency of alleles and genotypes will remain constant from ancient generations to current generations. In the HWE model, I considered, a single locus with two alleles, noted as 'A' (the reference allele) and 'B' (the derived allele) with allelic frequencies p and q respectively, and the frequency (A) = p and the frequency (B) = q where $p+q = 1$ denotes the frequency of the single locus with two alleles. The Equation $p^2+2pq+q^2=1$ describes the frequency of the three possible genotypes of the locus. Then, taking into account that the alleles of the controls in the dataset were in HWE, the frequency (AA) = p^2 for the homozygous wild type in the population, the frequency (BB) = q^2 for the homozygous carriers, and frequency (AB) = $2pq$ for the heterozygotes. A strong selective force on a genetic locus, such as the pressure of malaria, can distort the expected distributions under HWE, thus, in the attempt to discriminate between poor genotype call rate and selection due to malaria, assays were excluded if the HWE probability exceeded the 0.1% significance threshold ($P<0.001$) rather than the commonly used 5% ($P<0.05$) threshold in the control group.

4. Gender curation was achieved by "genotyping" three fixed nucleotide differences between the two copies of the amelogenin gene found on the X and Y chromosomes, where male samples were expected to be heterozygote for the alleles, and females are expected to be monomorphic. This process

allowed the curation of clinical gender (the reported gender) obtained through the KHDSS and the genetic gender inferred from the typed sample from the amelogenin variants.

All of these QC measures were undertaken to retain high quality, high information content data but came at the expense of decreased sample size and SNP coverage. Finally, prior to analysis, genotype data were recoded based on the human genome reference, any remaining heterozygous genotypes found in males for the X-chromosome SNPs were coded as missing, and samples for which gender or phenotype (case-control status) were undefined were removed.

2.2.2 The Genome-wide association study (GWAS)

The use of a large-scale dataset such as GWAS is undoubtedly of great advantage as the genetic basis of resistance to malaria is expected to be around many different malaria genes, individually contributing to a small population effect, which might be missed due to methodological issues such as sample size. We contributed human DNA samples and clinical data from 1,500 children who had been admitted for SM as defined in section 2.2.1 along with healthy controls, from the candidate gene case-control study discussed above to the MalariaGEN consortium for inclusion in a GWAS. I shall refer to this as the 'Kenya GWAS data'. Details about these data can be found in the original

study [158] where the primary aim of the experiment was to identify and investigate genetic loci that correlate with SM, and to investigate changes to standard methodology (including QC, imputation and association techniques) that are required to study data from African population.

2.2.2.1 Genotyping

Samples meeting DNA concentration and genotyping criteria (as described in section 2.2.1.3) with appropriate clinical data were selected for GWAS. All samples passing initial QC measures (as described in section 2.2.1.3) were sent to the Wellcome Trust Sanger Institute in Cambridge, UK for genotyping. Pre-genotyping QC was undertaken to confirm DNA concentration, genetic gender plus a set of 30 SNPs were generated on Sequenom® MassARRAY iPLEX™ platform to serve as a simple check on 'genotype ability' for the samples and to generate a set of 'barcode' markers Samples passing the previous steps were genotyped using the Illumina HumanOmni2.5-4 genotyping chip.

2.2.2.2 Quality control

A variety of QC steps were undertaken on the raw data as detailed in Table 2.2. A SNP calling threshold of >95% was applied to each SNP. Raw intensities were processed via software callers using different algorithms to assess cluster plots of signal intensities for each SNP. Consensus calls were made across the callers. SNPs less than the calling threshold were further removed.

SNP were further filtered based on minor allele frequency, and HWE P value. A minor allele threshold of >1% was applied to each SNP. An allele whose frequency was lower than the threshold was removed. A χ^2 value of lower than 10^{-4} was considered a strong evidence of deviation from HWE.

All samples with missingness >2.5% were excluded. Differential missingness between cases and controls was also examined, and related samples were removed by computing a pair–wise concordance between samples using a thinned set of approximately 178,775 SNPs chosen to be at least 0.005cM apart. Duplicated samples were also excluded. Ethnicity and gender information were compared with the genetic information for consistency and any population substructures detected in the data were removed.

2.2.2.3 Imputation

Even though a SNP passed the calling threshold, there were some samples for which the genotypes of this SNP were not successfully determined. In such a situation, the genotype data was phased using SHAPEITv2.rs644 [201], specifying 200 hidden states and an effective population of 17,469 as per recommendation for African populations in the software's documentation. Phased genotypes were imputed using IMPUTE v2.3.0 [202] an imputation tool.

Table 2.2. Quality control for the GWAS.

SNP filtering	
Excluded SNPs with:	
1.	Genotype missing rates >0.05
2.	Hardy-Weinberg P values $<1\times10^{-4}$
3.	Minor allele frequencies <0.01
Sample filtering	
Excluded samples:	
1.	Missing genotype rate>0.025
2.	One sample from each pair of related samples
3.	Population outliers detected using Principal Component Analysis
4.	Samples with inconsistencies between reported gender and genotype-determined gender
5.	Outliers for heterozygosity

2.3 Statistical analysis

To answer the thesis research questions, different statistical approaches were applied on separate datasets discussed in section 2.2. These were: (1) Single marker association analysis, (2) Haplotype structure and LD patterns, and (3) Multilocus effect (SNP-SNP interactions).

2.3.1 Single marker association analysis

In the candidate gene case-control association study, logistic regression models were used to test for associations, and estimate odds ratios and confidence intervals. Logistic regression is an adaptation of linear regression in which a logarithmic transformation "logit" is used to allow the analysis of a binary phenotype (i.e. a case or control). In this approach I modelled the SNP of interest assuming several related genetic models (additive, dominant, recessive and heterozygous advantage models) and reported the minimum p-values derived from these correlated models.

1. The "additive model": I assumed that the three genotypes of any given SNP had an independent effect. In other words, a model of dominance in which the heterozygote was at an immediate risk between the two homozygotes.
2. The "dominance model": individuals with a single copy of a disease susceptibility allele were at the same risk of contracting the disease as those

with two copies (i.e the presence of one copy of an allele increased or decreased the disease manifestation).

3. The "recessive model": individuals with only one copy of the susceptibility allele were at the same risk as those with none.
4. The "heterozygote model": individuals, who carry two different alleles, were compared to the other individuals.

Further explanations about the genotypic models of inheritance are summarised in Table 2.3.

First I fitted a logistic regression model as shown in **Equation 2.1** to predict disease association from each SNP individually. In the model I coded the SNP by two indicator variables:

a) One for the heterozygous carrier of the derived allele (X_1) and

b) One for the homozygous carrier of the derived allele (X_2)

In other words, I converted AA into "$X_1=0$, $X_2=0$", AB into "$X_1=1$, $X_2=0$", and BB into "$X_1=1$, $X_2=1$" where 'A' stands for the reference allele and 'B' stands for the derived allele. The case-control status response variable Y was coded as 1 for cases and 0 for controls.

$$log\frac{P(Y=1|X_1,X_2)}{1-P(Y=1|X_1,X_2)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \textit{(Equation 2.1)}$$

Secondly, I hypothesized that the SNP was not associated with the response variable for the null hypothesis (H_0), namely β_1 and β_2 were zero meaning that for the alternative hypothesis (H_1) either β_1 or β_2 were non-zero.

To measure the genetic effect, the standard odds ratio (ORs), defined as the odds of exposure among the cases divided by the odds of exposure among the controls, was used. With genotypes coded with respect to the alternate allele, and the phenotype coded with respect to the disease status then; an OR of 1 implied that exposed individuals are at no increased risk compared with those who are not exposed, and that an OR of >1 implied an increased risk (or positive association), while an OR <1 implied a protective effect (or negative association). Odds ratio were estimated by exponentiating the beta coefficients of the model. The above analysis was performed in all SM cases and within sub-clinical phenotypes of SM: CM, SMA, and RD.

Table 2.3. Coding of alleles for logistic regression analysis with respect to different models of inheritance.

MODEL	Ancestral Homozygotes	Heterozygotes	Derived Homozygotes
Autosomal chromosomes and female X chromosomes			
GENOTYPE	**AA**	**AB**	**BB**
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Heterozygote	0	1	0
Male X chromosomes			
GENOTYPE	**A**		**B**
Additive	0		2
Dominant	0		1
Recessive	0		1
Heterozygote	n.a.		n.a.

In the additive models ancestral-allele homozygotes were coded as 0 for both chromosomes autosomes and sex chromosomes, heterozygotes were coded as 1 and derived allele homozygotes were coded as 2 (including the male derived-allele homozygotes so that they are treated equivalent to the female X chromosome derived-allele homozygotes in the analysis). For all other models, the genotypes were coded as 0 or 1 depending on the model grouping requirements having 1 with respect to the model and derived allele. n.a.= not applicable.

2.3.1.1 Variance explained analysis

Estimation of the proportion of genetic variance explained by host genetic factors plays a significant role in understanding a complex trait. Often, the coefficient of determination (R^2) is used to account for this proportion. In my analysis, before calculating the R^2, I examined the LD patterns between the markers using standardized summary statistic r^2 [203]. Due to high LD in some of the SNPs within a gene, I only selected the most significant SNP for estimating the genetic variance where more than one SNP was typed within any single gene.

The proportion of variance explained by a single gene was estimated by including one SNP within each gene in a regression model with all covariates used in the single-SNP association analysis and, subsequently calculated the R^2 for the full model. I then subtracted the variance explained by a basic model in which only covariates were included from the variance obtained from the full model.

2.3.2 Multilocus effects (SNP-SNP interactions)

One disadvantage of considering only a single SNP effect is that it potentially neglects the joint effect of multiple SNPs, where some variants may have little marginal effect, but the effect of the variant is more evident when it is altered or

highlighted by another variant or variants. Therefore, a superior approach to the above single-SNP association analysis is the multilocus effect test which examines the association of a phenotype with multiple SNPs simultaneously.

In order to detect two-locus epistatic interactions in my data set, three representative algorithms were used: **PLINK** (see section 2.3.2.1) and ant-colony optimization algorithm (**AntEpiSeeker;** see section 2.3.2.2), along with my own novel method known as 'Single Nucleotide Polymorphism epistasis' (**SNPepistasis**; see section 2.3.2.3). The three computational methods were selected because of the clear-cut advantages given in detail in **Appendix A.1 Table A.1**. Table 2.4 further shows similarities and differences among the three methods.

Table 2.3. Similarities and differences among the three computational methods selected for epistasis detection

Attribute	PLINK	AntEpiSeeker	SNPepistasis
Exhaustive Search	Yes	No	Yes
Logit Model Assumed	No	No	No
Multi-Stage	No	Yes	No
Permutation Test	No	No	Yes
Bonferroni Correction	Yes	Yes	Yes
Adjusting of covariates	Yes	Yes	Yes
Applicability to case-control study design	Yes	Yes	Yes
Dimensionality scalability	Yes	Yes	Yes
Type of Test	Logistic regression	*Chi*-square	Logistic regression
Programming Language	C++	C++	R
Open source	Yes	Yes	Yes

2.3.2.1 PLINK

2.3.2.1.1 Introduction

PLINK is an open-source C++ single-threaded, command line program, described in detail elsewhere [204, 205]. It has been developed by Shuan Purcell at the Centre for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH) and the Broad Institute of MIT and Harvard, with the support of others [205]. The main purpose of PLINK is manipulating and analysing of large whole-genome datasets in their entirety. Its initial functionality covers five main domains: data management, summary statistics, population stratification, association analysis and identity-by-descent. It has a set of tools for analysis of the association of the complete genome. It was designed to perform a range of basic, large-scale analyses in a computationally efficient manner. In addition to its other functions, it can be used to investigate statistical epistasis.

Currently, PLINK provides both exhaustive and non-exhaustive search methods for detecting epistasis. The search for association of 2 loci with a phenotype is performed by a module called "epistasis". PLINK applies a logistic regression test for the evaluation of a locus × locus applied according to **Equation 2.2**. In **Equation 2.2**, p is the probability of having the disease, β_0 represents the null effect, β_1 and β_2 represent the main effects of each locus on the phenotype, and

β_3 represents the interaction term. The variables x_1 and x_2 contain information about the genotype in both loci and can be encoded in different ways depending on the genetic model of inheritance (see Table 2.3). The interaction term is encoded by $(x_1 * x_2)$.

$$\text{logit}(p) = \ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) \quad \textit{(Equation 2.2)}$$

The coefficients are estimated for each SNP as well as the interaction between them.

2.3.2.1.2 Algorithm

a) For each pair of loci x_1 and x_2, PLINK evaluates the model of interaction according to equation 2.2

b) Estimate overall Odds Ratio.

c) Apply the test Z-score (**Equation 2.3**).

The test Z-score is applied according to **Equation 2.3** where the variance V is estimated as: $\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}$

$$\text{Test } z-\text{score} = \frac{[OR(\text{cases} - OR(\text{controls})]_2}{v(\text{cases}) + v(\text{controls})} \quad \textit{(Equation 2.3)}$$

d) Carry out the control of multiple tests by applying Bonferroni correction (see section 2.4.4.1).

2.3.2.1.3 Parameter settings

Before using PLINK, the raw data were converted to a default file format .ped and .map. Typically PLINK requires two or more data files to perform desired operation: *.ped* and *.map* file in text format. The *.ped* file is a white-space delimited file, containing values of genotypes and information relating to the samples. It is made up of six mandatory and two optional columns. The mandatory columns contain *family ID, individual ID, paternal ID, maternal ID, sex* and *phenotype*, respectively. The two additional columns contain names of alleles. The *.map* file, as the name suggest contains the genetic map and has four columns: *chromosome number*, *SNP identifier*, it's *genetic distance* (in Morgans – unit for measuring the genetic linkage) and base-*pair position* (one kilobase equals to 1000 base pairs of DNA). Finally, upon uploading the two standard files (*.ped* and *.map*) into PLINK, the software performed some extra sanity checks on the input data, and later reformatted them to a binary format for its own use. The binary file format splits information into three files: the binary *.bed* file (containing genotype information), the text file *.fam* (storing phenotype information) and text file *.bim*, which is an extended *map* file (contains two extra columns with allele name).

To display an epistatic result, the interaction P-value was set larger than the default value in PLINK . The analysis was performed using the '--epistasis' command for all-SNPs × all-SNPs interactions and the total number of valid tests was saved into an output file containing six columns: chromosome of first SNP (CHR1), identifier for the first SNP (SNP1), chromosome of second SNP (CHR2), identifier for the second SNP (SNP2), χ^2 statistics (STAT) and an interaction p-value (P).

2.3.2.2 AntEpiSeeker

2.3.2.2.1 Introduction

AntEpiSeeker is an algorithm developed in C++ for detecting epistatic interactions in large-scale case-control studies, using an ant-colony optimization (ACO) algorithm derived from Dorigo and Gambardella *et. al* [206]. The use of χ^2 values as score function to measure the association between a SNP set and the phenotype has been made computationally efficient.

2.3.2.2.2 Algorithm

AntEpiSeeker is a two-step algorithm. In the first step, it uses a χ^2test for an association for the interaction between two-pairing loci and the phenotype, and initially no assumption about the interaction is made in AntEpiSeeker. In the second step, AntEpiSeeker conducts a thorough, exhaustive search of epistatic interactions within the set of SNPs that show some evidence of interactions and

within the reduced set of SNPs with top ranking weights (pheromone levels). The use of SNP sets showing some evidence of interaction (much smaller than the available SNPs in the data) enhances the power of detecting pure epistasis based on greatly reducing the computation time involved and the SNP set with top ranking weights is used to detect epistasis among the SNPs with big marginal effects. A detailed description about the AntEpiSeeker algorithm can be found in [207, 208].

2.3.2.2.3 Parameter settings

Before running AntEpiSeeker, the genotype data file was converted to a comma-delimited file, with the first row specifying the SNP names and all subsequent rows contained SNP data for each sample. The last column indicated the case-control status (0 indicating a control and 1 indicating a case). A separate parameter file named "parameters.txt" specified the parameters needed to run the program. For my two-locus interaction model the parameters were set as per Table 2.4. After running the epistatic interaction, three output files were generated "AntEpiSeeker.log" that contained some intermediate results, "results_maximized.txt" that contained all detected epistatic interactions, and the 'user-specified output' file showing the epistatic interactions with minimized false positives. The user specified output file included the following columns: the locus name, χ^2 test value and a p-value.

Table 2.4. Parameters settings for the AntEpiSeeker algorithm.

Parameter	Settings	Parameter description
iAntCount	114	Number of SNPS
iItCountLarge	11	Number of iterations for the large haplotypes
iItCountSmall	58	Number of iterations for the small haplotypes
alpha	1	Weight given to an interaction
iTopModel	1000	Number of top ranking haplotypes in the first stage
iTopLoci	200	Number of loci with top ranking interactions in the first stage
phe	100	Initial number of interactions for each locus
largehapsize	6	Size of the large haplotypes
smallhapsize	3	Size of the small haplotypes
iEpiModel	2	Number of SNPS in an epistatic interaction
pvalue	0.05	P-value threshold (after Bonferroni correction)
INPFILE	Kenya.txt	Input file name for case-control genotype data
OUTFILE	Kenya_result.txt	Output file name for detected epistatic interactions

2.3.2.3 SNPepistasis

2.3.2.3.1 Introduction

SNPepistasis is my proposed novel exhaustive method for detecting two-locus epistatic interactions in a case-control study. The algorithm is implemented in the R environment which is a free open source statistical software (http://www.r-project.org).

2.3.2.3.2 Algorithm

I have adopted an exhaustive approach that examines all possible interactions among the SNPs. This ensures that I do not miss any statistically significant interactions. My method can easily be extended to deal with heuristic search approaches. In order to determine significant epistatic interactions for two loci, I have used a logistic regression model whereby I have considered two models; a model that includes an interaction term (saturated model) and a model that omits an interaction term (reduced model) such as the following:

Without interaction:

$$log\frac{P(Y=1)}{P(Y=0)} = \mu + \alpha X_1 + \beta X_2 \qquad \textit{(Equation 2.4)}$$

With interaction:

$$log\frac{P(Y=1)}{P(Y=0)} = \mu + \alpha X_1 + \beta X_2 + \gamma X_1 X_2 \qquad \textit{(Equation 2.5)}$$

Where the response item at the left hand represents the log odds of the disease risk, the right hand represents the independent effect caused by the two SNPs, X_1 and X_2, and the multiplicative term (X_1,X_2) represents the epistatic interaction. The terms: α, β, γ are the size of the effects which are determined in the regression analysis and range from 0–1. The epistatic interaction between a pair of SNPs and the phenotype was inferred by testing whether the regression coefficient γ was equal to zero, and this was done by using a likelihood ratio test (LRT), which expressed how many times more likely the data are under one model (i.e. with interaction term) than the other (i.e. without interaction term).

A P-value as a measurement of the confidence of the epistatic interaction was then computed. Considering that the theoretical P value derived from LRT may not reflect the true null distribution, I further used permutation test to account for multiple comparisons. Finally, an output file "SNPepistasis_output.csv" that contained all detected epistatic interactions was generated. The pseudocode for SNPespistasis method is shown in **Appendix B, SNPepistasis pseudocode B.1.** Finally, Quantile-Quantile plots (Q-Q plots) were generated using customised scripts implemented in R to detect inflation of statistics due to population stratification.

2.3.3 Haplotype structure and LD patterns

Haplotype-based methods utilising SNPs offer a more powerful approach to complex disease gene mapping, because they capture the association between causal mutations and the ancestral haplotypes on which they arose. For the current study, the haplotype patterns across the HbS and α^+thalassaemia loci were constructed and pairwise LD estimated for the surrounding SNPs in each gene using standardized summary statistics [203] D' and r^2. These calculations were performed using customised R scripts as described in Chapter 5.

2.3.4 Multiple testing corrections

Multiple testing is the major concern for the SNP-by-SNP search algorithm. Without proper adjustment, it is likely that false positive conclusions will arise (i.e. Type I error). The frequentist method for controlling for false positive associations is by controlling the significance level, α. The usual choice of α is 0.05, which implies that the probability of being false positive in all the tests carried out is less than 5%. Here I highlight two commonly used approaches for adjusting for the multiple-testing problem though various other methods have been proposed for controlling issues derived from multiple testing in association studies.

2.3.4.1 Bonferroni correction

Bonferroni correction is often used to control for Type I errors. If n SNPs are tested for association, the Bonferroni corrected for each test at significance level, α is α/n. For genotype data, the value of "n" can be substantially large and depends on the SNP chip used in GWAS studies (e.g. in my data the chip has 2.5 million SNPs). Bonferroni does not account for correlations between SNPs (LD) and therefore can be overly conservative particularly in areas of high LD.

2.3.4.2 Permutation testing approach

Permutation testing is a simulation based resampling method, which controls the issues of multiple testing by comparing observed p-values with p-values estimated by repeated perturbation of the data and evaluating how often the observed p-value can be obtained by chance [209]. In this study, I used the permutation approach to select significance levels for further analyses. I randomly generated 1,000 permutations of phenotypes across the subjects and repeated the analysis in order to find the maximum test value of each permutated phenotype and reported these maximum test values. Further details regarding the analytical methods are given in their respective chapters.

2.4 Ethical considerations

Approval for the recruitment of participants, collection of blood samples, DNA preparation, genotyping was provided by the KEMRI/Wellcome Trust Scientific Coordinating Committee (SCC) in Kilifi, and by the KEMRI Scientific Steering Committee (SSC number=1192) and the KEMRI/National Ethical Review Committee (ERC) in Nairobi and informed consent from the parents or guardians of each study participants.

Chapter 3

Association analysis between human candidate malaria-protective genes and malaria

Abstract

Host genetic factors are thought to be important modifiers of severe and fatal malaria, yet there is still a lack of knowledge about which genes are involved or the type or degree of protection that they might afford. With this in mind, I have undertaken a case–control study in a rural area on the Coast of Kenya to evaluate associations between SM and its sub-phenotypes with candidate malaria resistance genes. A total of 2245 SM cases and 3949 controls from Kilifi populations were typed for a total of 136 SNPs in 71 malarial candidate genes using the Sequenom® iPLEX platform and for α^{+}thalassaemia using conventional PCR. I found a number of RBC polymorphisms (*HBB*, *HBA*, *G6PD*, *FREM3* (surrogate for *GYP*), *INPP4B*, *ATP2B4*, *ABO*) being putatively associated with differential susceptibility to SM ($P<0.005$). The total variability in the risk of SM that could be explained by all the above malaria candidate

genes additively was 7.6%. Overall, this study represents the first association study from the Kilifi population involving a large number of host genetic factors with susceptibility or resistance to SM.

3.1 Introduction

The risk factors for severe disease and the reasons why some children die of the disease and not others are poorly understood. Nonetheless, it is known that host genetic factors, parasite genetic make-up, host age, state of immunity and genetic background play a significant role in determining an individual's susceptibility to many infectious diseases, including malaria.

While a vast body of evidence have been carried out on associations of candidate genes with resistance and susceptibility to human malaria [190, 210-212], most of these associations have not been conclusively replicated across multiple populations. Study design features (e.g. case definitions, sample size, study type) and allelic heterogeneity has been proposed as reasons [42]. There is a need to unravel host genetic factors that provide a significant contribution to the variability observed in malaria using a large-scale study which could lead to a more unbiased approach to genetic discovery [45, 158].

To begin to explore such contribution in this Chapter, I have utilised a large-scale well-characterised case-control study of SM to reappraise 136 published

SNPs, including loci related to haemoglobinopathies, cytokines and other immune mediators, thought to be involved in malaria pathogenesis, together with their receptors, and promoters, and in addition to SNPs that have shown early promise for associations in a GWAS undertaken by the MalariaGEN consortium [213]. I have further estimated the proportion of the total variation explained by these genes in a Kilifi population.

3.1.1 Objectives

1. Assembling a case-control resource for understanding of protective host genetic factors of SM in Kilifi.
2. To reappraise a range of candidate malaria-protective genes in a Kilifi population.
3. To estimate the proportion of the total variation in the risk of SM that can be explained by these candidate genes.

3.2 Material and methods

3.2.1 Study participants

A full description of the study design and study participants is presented in Chapter 2 (section 2.1 and 2.2 respectively). Severe malaria cases included individuals with CM, SMA, RD or other features of clinical severity who were

admitted to the high dependency ward of KCH between June 1995 and June 2008 while controls were children born consecutively within the same study area as cases who were recruited at 3-12 months of age into the Kilifi Genetic Birth Cohort Study between August 2006 and September 2010 as described in Chapter 2 (section 2.2.1.1).

3.2.2 DNA extraction and genotyping

Genomic DNA was extracted from samples of blood using methods described in Chapter 2 and samples were genotyped using the Sequenom® MassARRAY iPLEX™ platform (Agena Biosciences, Hamburg, Germany) for 136 SNPs, and PCR was used to detect deletions in the α^{+}thalassaemia gene (see Chapter 2 (section 2.2.1.2).

3.2.3 Statistical analysis

Single-locus association-analysis adjusted for gender, ethnic group and the HbS polymorphism, were performed comparing different phenotypes group of interest: a) all severe malaria vs healthy controls; b) all major non-mutually exclusive sub-phenotypes: CM, SMA, RD vs healthy controls. Genotypic deviations from the HWE were assessed using a χ^2 statistical test. SNPs were excluded from the analysis if there was at least 10% of the genotype calls missing or a significant

deviation from the HWE ($p<0.0001$) in controls, or if the overall minor allele frequency was less than 0.01 (see Chapter 2, section 2.3.2.3).

Logistic regression models were used to test for associations, and estimate the odds ratio and 95% confidence intervals. I modelled the SNP of interest assuming several related genotypic models of inheritance (additive, dominant, recessive, heterozygous advantage models) and reported the minimum p-value from these correlated models. Further explanations about the tests are described in Chapter 2 (section 2.3.1).

Statistical analysis of SNPs on the X-chromosome was performed for each gender separately. Performing multiple statistical tests leads to inflation of false positives. As a result, I used a permutation approach that accounted for the correlation between markers and tests, to determine a p-value cut-off for statistical significance of 0.005 (Chapter 2, section 2.3.4.2). Further analysis was performed to evaluate the proportion of variance explained by a single gene or any given locus by including the gene in a regression model with all covariates used in the association analysis and calculating the R^2 coefficient for the full model (Chapter 2, section 2.3.1.1).

3.3 Results

3.3.1 Overview of study population and exclusion criteria

The process of case ascertainment is summarised in Figure 3.1. In summary, there were 62,281 admissions to the KCH paediatric wards during the study period (June 1995-June 2008) among which 16,174 (26%) had a positive malaria slide and 2245 (3.6%) fulfilled the criteria for SM. The proportion of malaria-related admissions relative to admissions due to all causes by year is summarised in Figure 3.2.

3.3.1.1 Summary of Severe malaria cases

The baseline characteristics of the study population are summarised in Table 3.1. Of the 2245 cases, 1234 (55%) had CM, 687 (31%) had SMA, 687 (31%) had RD, 670 (30%) had CM alone, 230 (10%) had SMA alone, 157 (7%) had RD alone, and 436 (20%) with other signs of SM. The male : female ratio was 1.06: 1. The median age of the SM cases was 28 months with an interquartile range (IQR) of 15-43 months. The median age of children with CM, SMA and RD was 28 months, 20 months and 25 months respectively. The SM cases were predominantly of Giriama 1326 (59%) and Chonyi 527 (23%) ethnic origin.

The overlap between SM phenotypes and their case-fatality rates are summarised in Figure 3.3. Of 2245 cases, 263 died (56% male), giving an overall case fatality rate of 12%. The median age of the cases that died was 28 months (IQR: 2–132 months). Children with CM had the highest case fatality rate ranging from 11% for CM accompanied with SMA to 23% when CM was accompanied by both SMA and RD. RD alone had a case fatality rate of 8%. The lowest case fatality rate was observed in children admitted with SMA as the only clinical feature 7%.

3.3.1.2 Summary of healthy controls

A total of 34,714 children were identified between May 2006 and February 2011 as being eligible for inclusion in the ongoing prospective Kilifi Genetic Birth Cohort Study (see Chapter 2 section 2.1.1). 16,275 healthy children aged 3–12 months were consented and enrolled in this study. Of these, 11,050 (68%) have so far been genotyped for various malaria candidate polymorphism and of these, 3,949 were selected using pseudo-random sampling as potential controls to be included in the final case-control analysis (Figure 3.4). The selected controls had a median age of 6 months, were well balanced for gender (50% were male) and were from predominantly of Giriama 1836 (46%) and Chonyi 1421 (36%) ethnic origin (Table 3.1). No significant difference was observed between the cases and controls with regard to ethnicity.

Figure 3.1. Schematic process for the recruitment of cases into the candidate gene case-control study.

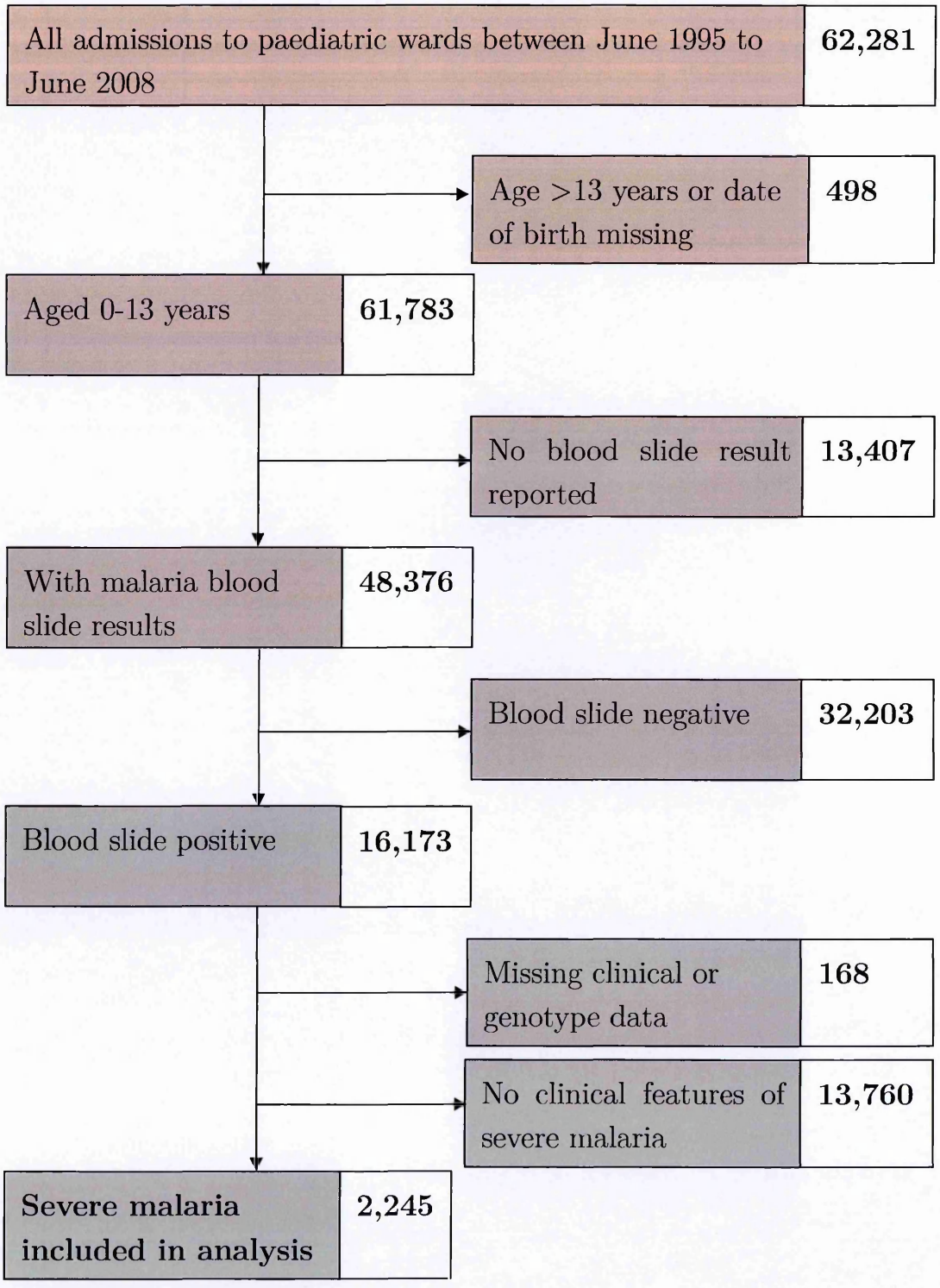
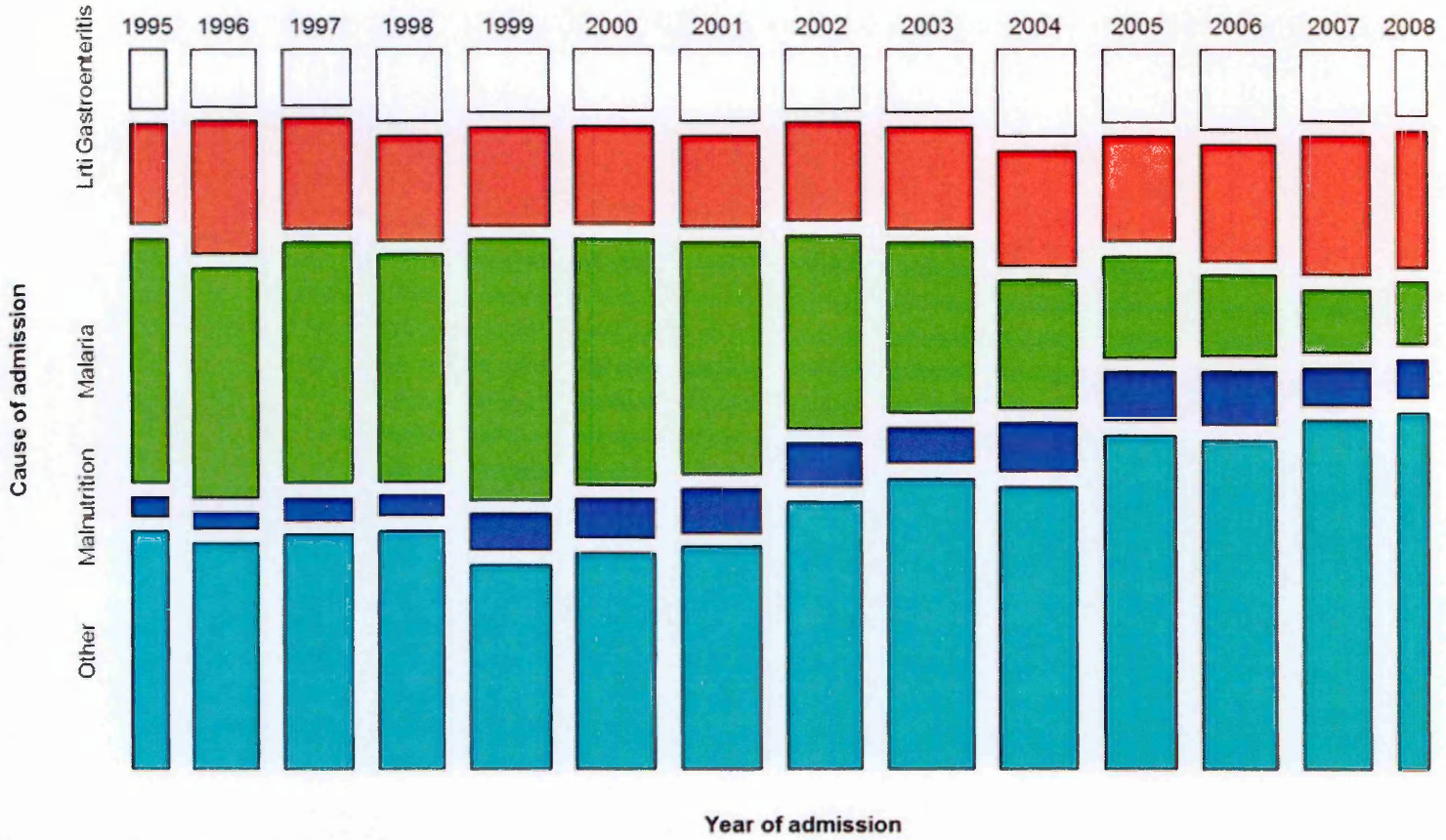


Figure 3.2. Top five causes of hospital admissions at KCH between June 1995–June 2008.



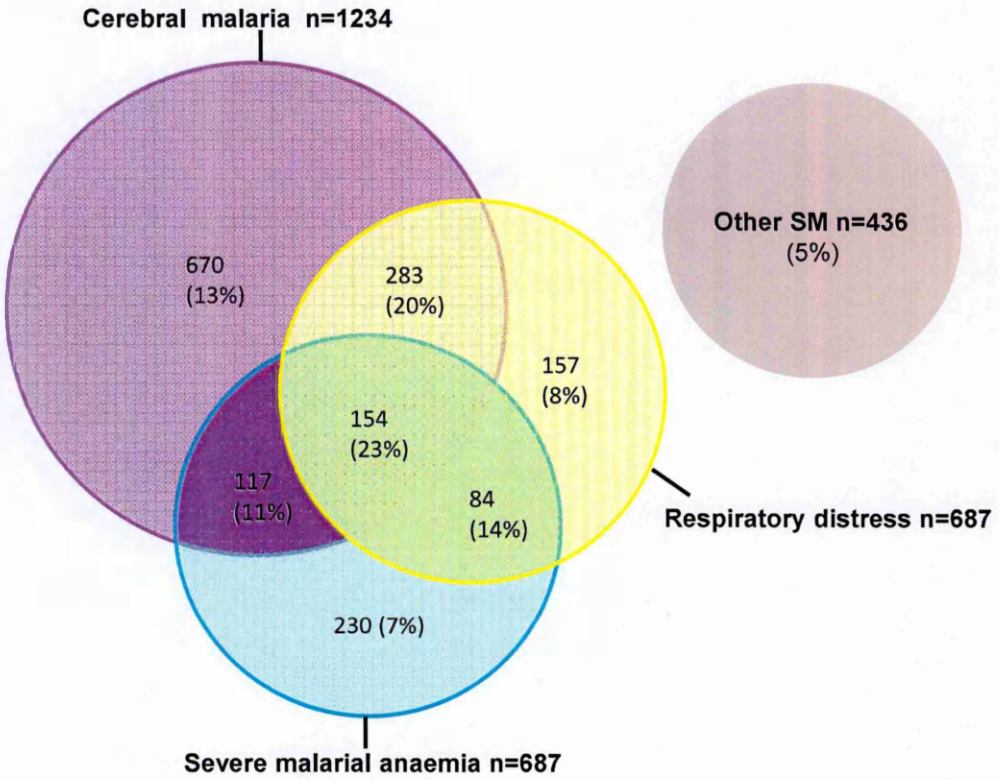
The cause of hospital admission is shown by year from left to right. Colours indicate the cause of admission indicated on the left axis. From the figure, we can see the trend of malaria admissions overtime (the green bars). Abbreviations: Lrti= Lower respiratory tract infection.

Table 3.1. Baseline and clinical characteristics of the candidate gene case-control study.

Characteristics		Cases								Controls
		All SM	Clinical phenotypes							
			All CM	All SMA	All RD	CM only	SMA only	RD only	Other SM	
All subjects, No. (%)		2245 (100)	1234 (55)	687 (31)	687 (31)	670 (31)	230 (10%)	157 (7%)	436 (19)	3949
Age, months, *median (range)*		28 (2-155)	28 (2-155)	19 (4-132)	25 (2-123)	31 (2-155)	19 (2-132)	28 (4-123)	32 (4-155)	6 (3-12)
Gender, *No. male (%)*		1157 (52)	628 (51)	358 (52)	358 (52)	337 (50)	134 (58)	76 (48)	226 (52)	1992 (50)
Ethnicity, *No. (%)*	Giriama	1321 (59)	744 (60)	386 (56)	380 (55)	412 (61)	135 (59)	83 (53)	268 (62)	1335 (46)
	Chonyi	527 (23)	264 (21)	176 (26)	181 (26)	140 (21)	62 (27)	50 (32)	97 (22)	1411 (36)
	Kauma	171 (8)	98 (8)	48 (7)	64 (9)	49 (7)	16 (7)	17 (11)	25 (6)	440 (11)
	Others	226 (10)	128 (10)	77 (11)	62 (9)	69 (10)	17 (7)	7 (5)	46 (10)	273 (7)
Parasitaemia, $\log_{10}$ value, µL		10.5 (2.5)	10.4 (2.6)	10.4 (2.3)	10.9 (2.5)	10.2 (2.6)	10.5 (2.3)	11.0 (2.6)	10.4 (2.6)	-
MCV, fL		74.1 (9.4)	74.3 (9.3)	75.4 (10.6)	74.8 (9.8)	73.7 (8.7)	74.9 (10.2)	73.0 (8.4)	73.3 (8.9)	-
Platelets, $\times 10^6$/ L		163 (146)	170 (145)	139 (118)	154 (147)	186 (149)	133 (103)	159 (165)	168 (164)	-
Haemoglobin, g/dL		6.6 (2.5)	7.0 (2.4)	3.8 (1.2)	6.3 (2.5)	8.0 (2.0)	3.8 (1.2)	7.6 (1.9)	7.9 (2.0)	-

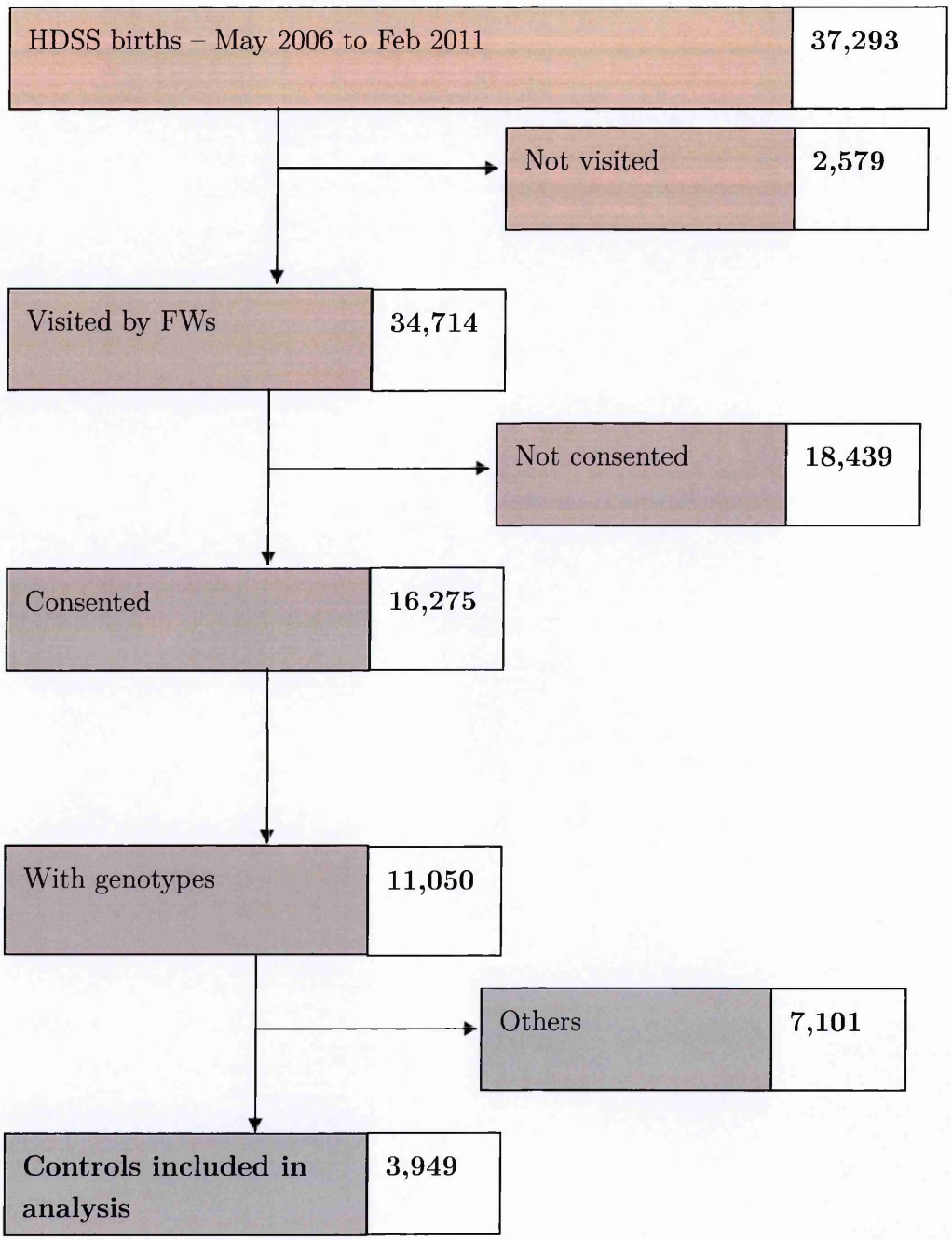
Numbers (with percentages) are presented for categorical variables. Means are presented with standard deviation for continuous variables, unless otherwise stated. The severe malaria cases (All SM) are further grouped into: All CM =cerebral malaria (Blantyre Coma Score ≤ 2), All SMA = severe malarial anaemia (HB <5 or haematocrit <15%), All RD = respiratory distress (deep breathing= yes), CM only=cerebral malaria as only clinical feature, SMA only=severe malarial anaemia as only clinical feature, RD only=respiratory distress as only clinical feature and Other SM= severe malaria cases without any of the major sub-phenotypes .

Figure 3.3. Venn diagram showing the overlap of the three major sub-phenotypes of SM and the subset of children with other severe syndromes in the candidate gene case-control study.



Total numbers (N) are shown with the case fatality rate of each sub-phenotype shown in parenthesis. Other SM=severe malaria cases without any of the major sub-phenotypes. Children may appear in more than one category.

Figure 3.4. Selection of healthy controls for the candidate gene case-control study.



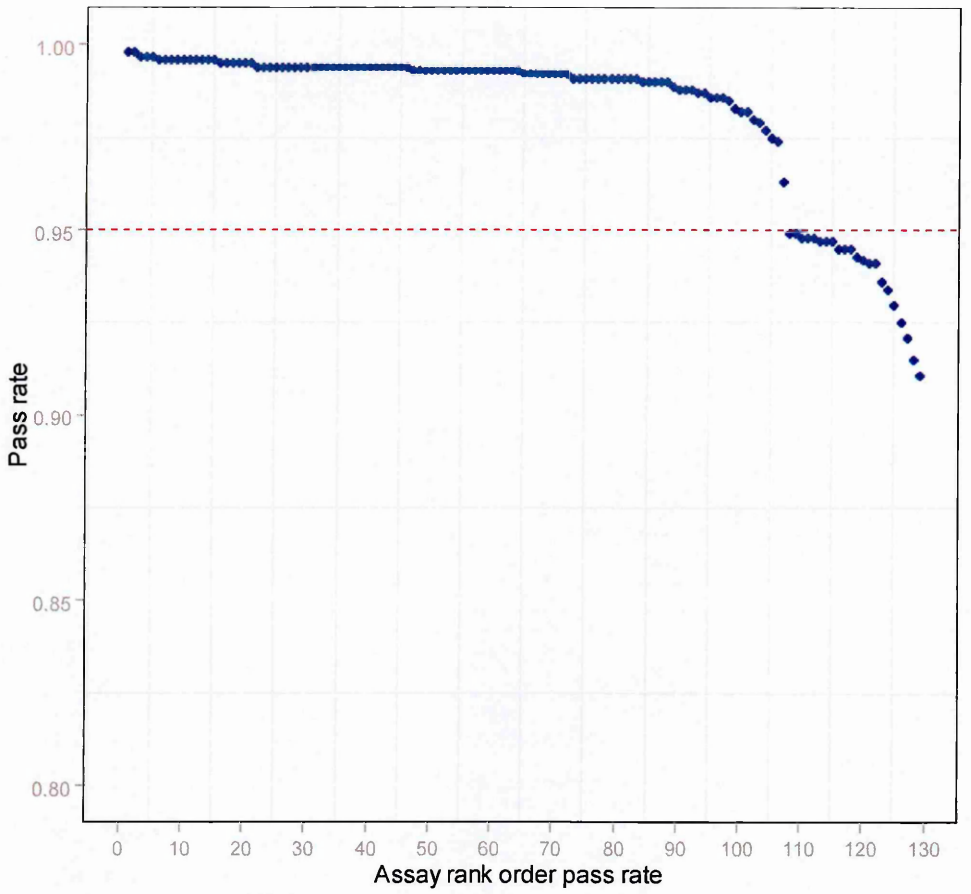
3.3.2 Assay pass rate

Genotype call rates for each multiplex were checked and assays with call rates of <95% were excluded from further analyses. A total of 136 assays were performed, with 121 assays achieving a pass rate of ≥ 95% (Figure 3.5).

3.3.3 Sample pass rate

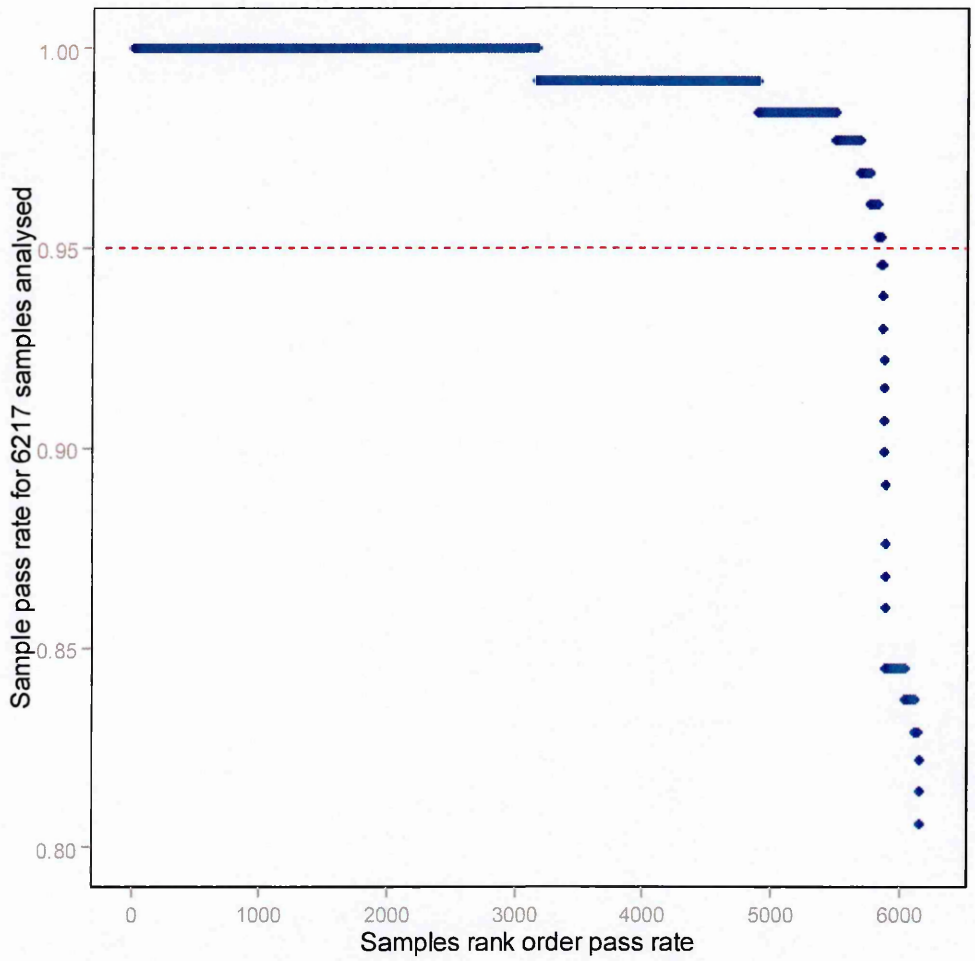
The performance of sample collection across the 136 SNPs was calculated as sample pass rate. Samples with call rates of <95% were excluded from further analyses. A total of 6,214 samples were assayed, with 6,194 samples achieving a pass rate of ≥ 95% (Figure 3.6).

Figure 3.5. The performance of 136 assays successfully genotyped in the candidate gene study.



Assays have been ranked in order of pass rate from high to low (x-axis) and plotted against their pass rate (y-axis). A total of 136 assays were performed, with 121 assays achieving a pass rate of ≥95%.

Figure 3.6. Performance of the samples included in the candidate gene case-control study.



Samples have been ranked in order of pass rate from high to low (x axis) and plotted against their pass rate (y axis). A total of 6,217 samples were assayed, with 6,194 samples achieving a pass rate of ≥95%.

3.3.4 Single-SNP association analysis

Overall, fifteen SNPs were removed from the analysis because they were monomorphic (*BCAS3*-rs184142841, *GRIP1*-rs192909543, *GUSBP5*-rs148111931, *HBB*-rs33950507, *HBB*-rs33930165, *ICAM1*-rs1799969, *LTBP2*-rs74063230, *OXNAD1*-kgp9483807, *OXNAD1*-rs75180423, *OXNAD1*-rs79691057, *PLEKHG1*-rs14422409, *RHOG*-rs138826089), had high rates of missing genotype calls (*GYPB*-rs191338817, *TNF*-rs1800750), or deviated from HWE among controls (*RPS6KL1*-rs3742785), leaving 121 polymorphic SNPs that could be analysed for their association with the four SM groups.

Figure 3.7 shows the minimum p-values from the genotypic tests applied to the autosomal SNPs. Twenty one loci including α^{+}thalassaemia were found to be highly associated with SM ($P \leq 0.005$). Interestingly, majority of these associations were related to the RBC polymorphisms. The most significant association being observed with *HBB*-rs334 located on chromosome 11 ($P=1.35\times10^{-59}$). The remaining significant loci, ranked by chromosome position were: four SNPs in the *ATP2B4* gene located on chromosome 1 (rs10900585, P=0.005; rs1541255, P=0.002; rs55868763, P=0.002), one SNP on the *IL10* gene located on chromosome 1 (rs1800890, P=0.002), three SNPs on the *LPHN2* gene located on chromosome 1 (rs4650365, P=0.004; rs72933304, P=0.002;

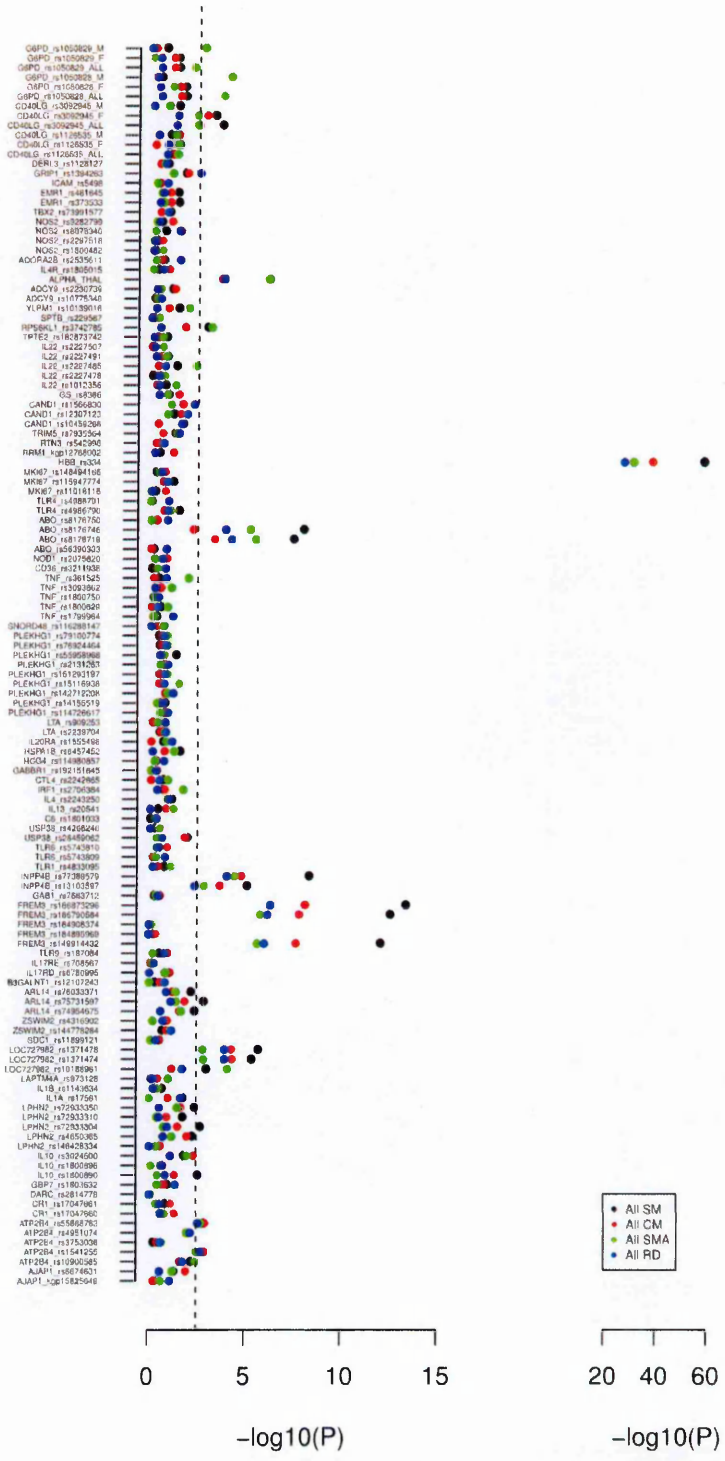
rs72933350, P=0.004), three SNPs on the *LOC727982* gene located on chromosome 2 (rs1371474, $P=5.68\times10^{-6}$; rs1371478, $P=2.57\times10^{-6}$), three SNPs on the *ARL14* gene located on chromosome 3 (rs74954675, P = 0.004; rs75731597, P=0.001), three SNPs on the *FREM3* gene located on chromosome 4 (rs149914432, $P=6.17\times10^{-13}$; rs186790584, $P=1.98\times10^{-13}$; rs186873296, $P=3.13\times10^{-14}$), two SNPs on the *INNP4B* gene located on chromosome 4 (rs13103597, $P=7.57\times10^{-6}$; rs77389579, $P=5.85\times10^{-9}$), two SNPs on the *ABO* gene located on chromosome 9 (rs8176719, $P=2.14\times10^{-8}$; rs8176746, $P=9.72\times10^{-9}$), one SNPs on the *RPS6KL1* gene located on chromosome 14 (rs3742785, P=0.001) and one other on the *HBA* gene located on chromosome 16 (α^{+}thalassaemia, $P=1.35\times10^{-6}$) (Figure 3.7 and Table 3.2).

Odds ratios for each of the four phenotypes are depicted using Forrest plots in Figure 3.8. The greatest magnitude of effect was seen in the *HBB* and *FREM3* loci (85% and 45% relative reduction in SM, respectively). In contrast to a number of previous studies, which have suggested that α^{+}thalassaemia might specifically protect against severe malaria anaemia [62, 66, 214], in this analysis it was associated with protection from all major phenotypes of SM (Figure 3.8; Table 3.2).

The rs1050828 (*G6PD*-202, X-chromosome) SNP was weakly associated with a decreased risk of SM overall (OR 0.82, 95% CI 0.70-0.97, P=0.01), but by contrast was associated with a paradoxically increased risk of SMA (OR 1.80, 95% CI 1.34-2.40, P=0.0001) (Table 3,3). Further interrogation of the data revealed that this was explained by the fact that the *G6PD*-202 deficient heterozygous females (OR 0.82, 95% CI 0.70-0.97, P=0.01) were relatively protected against SM, but not the more severely deficient hemizygous males (OR 1.10, 95% CI 0.92-01.33, P=0.29). The deficient hemizygous males were found to be susceptible to SMA (OR 1.86, 95% CI 1.36-2.54, P=0.0001) (Figure 3.7, Figure 3.8 and Table 3.3).

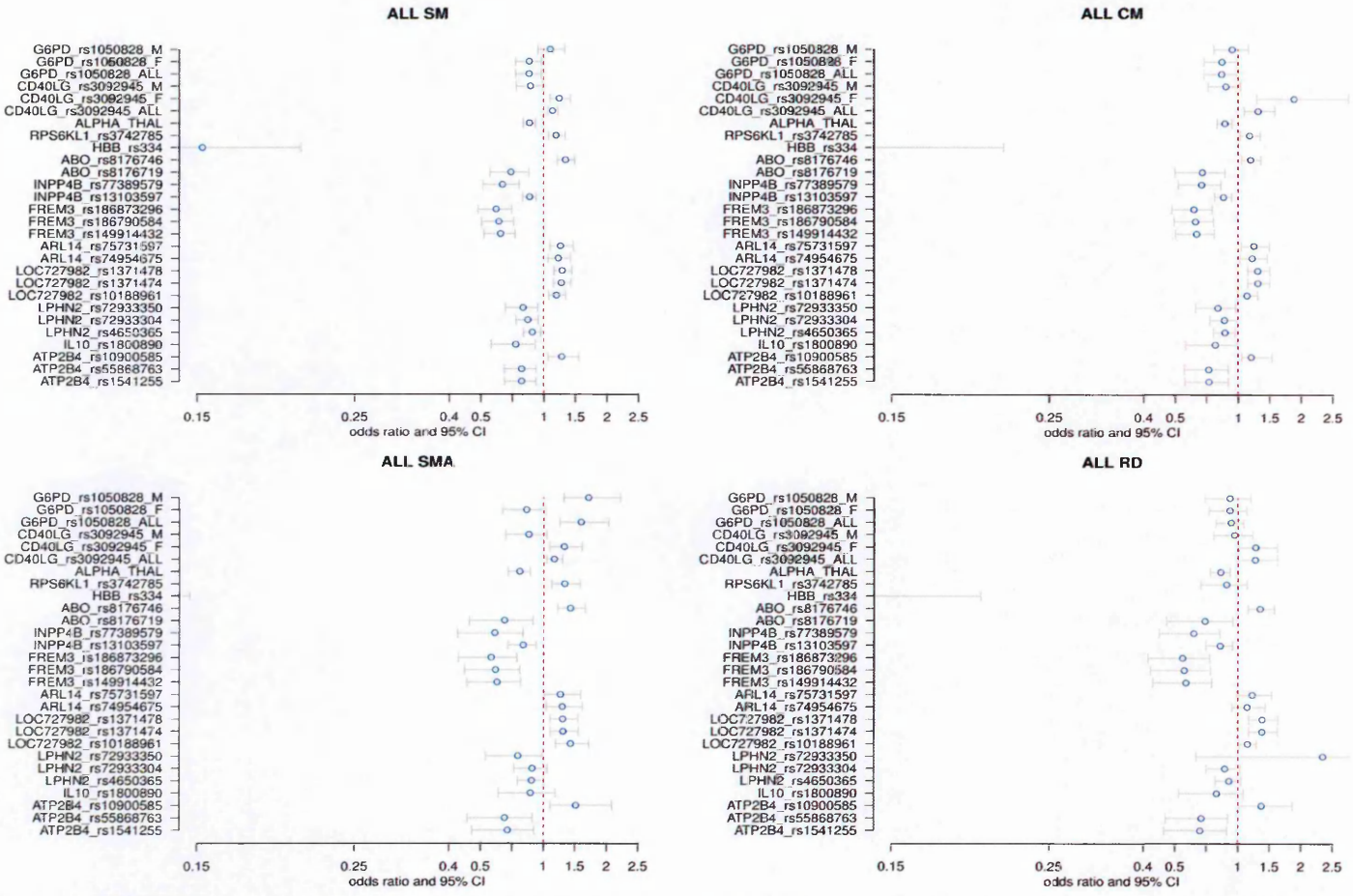
CD40L-rs3092945, a glycoprotein X-linked gene was found to be associated with an increased risk to SM (OR 1.14, 95% CI 1.06-1.22, P=0.0002) and SMA (OR 1.35, 95% CI 1.08-1.68, P=0.009). I observed an association between rs3092945-*CD40LG*-727 and both SM (OR 1.24, CI: 1.10-1.41; P=0.0006) and CM (OR 1.97, CI: 1.31-2.96; P=0.002) for females and a marginal association in males with SM as shown in Figure 3.7, Figure 3.8 and Table 3.3. Association analysis minimum p-values for mutually-exclusive SM phenotypes (i.e.CM only, SMA only and RD only) are presented in **Appendix C, Figure C.1.**

Figure 3.7. Shows the distribution of minimum p-values from the genotypic tests for the severe malaria and its major sub-phenotypes.



The x-axis shows the −log10 p-value and the y-axis shows the markers included in the analysis. The dashed line represents a p-value threshold of <0.005 as determined by permutation test (Chapter 2, section 2.3.4.2). Abbreviations see footnote, Table 3.1.

Figure 3.8. Shows Forrest plots for association with severe malaria and its major sub-phenotypes that were analysed.



The ORs and 95% CIs (grey bars) are shown for the markers that were found to be significantly associated with All SM. Loci are ordered by chromosome position. An OR= 1 for no effect is highlighted by the vertical dashed line.

Table 3.2. Autosomal SNPs with significant evidence of association with severe malaria and major sub-phenotypes.

Gene	SNPID	Alleles†	Chr	Frequency of Derived allele in Controls (Norm/Het/Hom)	Case phenotype	Frequency of Derived allele in Cases (Norm/Het/Hom)	MO	OR	LCI	UCI	P
ATP2B4	rs10900585	G/T	1	0.33(1644/1689/429)	All SM	0.33 (951/1042/207)	R	1.29	1.07	1.56	0.005
					All CM	0.32 (545/565/113)	R	1.21	1.06	1.54	0.02
					All SMA	0.30 (304/303/52)	R	1.52	1.11	2.09	0.003
					All RD	0.32 (301/321/158)	R	1.38	1.02	1.87	0.02
	rs1541255	A/G	1	0.33(1741/1774/418)	All SM	0.31 (1010/1016/185)	R	0.74	0.62	0.89	0.001
					All CM	0.30 (581/549/96)	R	0.69	0.54	0.87	0.001
					All SMA	0.29 (323/293/49)	R	0.69	0.47	0.87	0.003
					All RD	0.30 (323/313/49)	R	0.63	0.46	0.85	0.001
	rs55868763	G/C	1	0.33(1739/1778/414)	All SM	0.31 (1010/1016/183)	R	0.74	0.62	0.89	0.001
					All CM	0.30 (582/550/95)	R	0.68	0.54	0.87	0.001
					All SMA	0.29 (322/292/47)	R	0.62	0.45	0.85	0.001
					All RD	0.30 (321/313/49)	R	0.63	0.47	0.87	0.002
IL10	rs1800890	A/T	1	0.24(2270/1418/257)	All SM	0.23 (1313/821/199)	R	0.69	0.55	0.88	0.002
					All CM	0.23 (725/443/57)	R	0.74	0.55	0.99	0.03
					All SMA	0.24 (394/256/37)	R	0.83	0.58	1.19	0.3
					All RD	0.23 (400/252/133)	R	0.75	0.52	1.09	0.12
LPHN2	rs4650365	T/C	1	0.20(2482/1313/141)	All SM	0.19 (1468/664/83)	H	0.85	0.76	0.95	0.004
					All CM	0.19 (809/362/56)	H	0.83	0.72	0.96	0.009
					All SMA	0.18 (445/200/22)	H	0.85	0.71	1.01	0.05
					All RD	0.19 (447/210/27)	H	0.88	0.73	1.05	0.15
	rs72933304	C/A	1	0.08(3297/619/26)	All SM	0.07 (1912/296/8)	H	0.8	0.69	0.92	0.001
					All CM	0.07 (1057/168/5)	H	0.82	0.69	0.98	0.02
					All SMA	0.07 (570/93/3)	H	0.85	0.68	1.06	0.15
					All RD	0.07 (587/93/3)	H	0.83	0.66	1.03	0.09
	rs72933350	T/C	1	0.05(3569/364/7)	All SM	0.04 (2046/166/3)	H	0.75	0.62	0.92	0.004
					All CM	0.04 (1134/93/2)	H	0.76	0.6	0.96	0.02
					All SMA	0.04 (619/47/1)	H	0.72	0.52	0.98	0.03
					All RD	0.05 (624/57/3)	H	2.35	0.60	9.18	0.24

LOC727982	rs10188961	A/G	2	0.40(1418/1871/645)	All SM	0.42 (727/1096/388)	D	1.21	1.08	1.35	0.0009
					All CM	0.41 (420/607/198)	D	1.14	1.00	1.31	0.05
					All SMA	0.45 (193/349/123)	D	1.44	1.20	1.72	0.00007
					All RD	0.43 (222/336/124)	D	1.16	1.03	1.30	0.01
	rs1371474	T/C	2	0.27(2112/1496/325)	All SM	0.29 (1080/969/167)	H	1.29	1.16	1.43	4.19×10^{-6}
					All CM	0.29 (601/545/184)	H	1.32	1.16	1.50	0.00004
					All SMA	0.30 (316/297/54)	H	1.32	1.11	1.56	0.001
					All RD	0.29 (328/313/44)	H	1.39	1.18	1.63	0.0001
	rs1371478	C/T	2	0.26(2123/1466/320)	All SM	0.29 (1088/958/163)	H	1.30	1.17	1.45	1.90×10^{-6}
					All CM	0.29 (606/536/84)	H	1.32	1.15	1.50	4.61×10^{-5}
					All SMA	0.30 (318/291/54)	H	1.31	1.11	1.55	0.001
					All RD	0.29 (331/308/43)	H	1.39	1.18	1.64	0.0001
ARL14	rs74954675	A/C	3	0.08(3330/546/31)	All SM	0.10 (1794/391/15)	H	1.24	1.07	1.43	0.004
					All CM	0.09 (995/218/6)	H	1.23	1.03	1.46	0.02
					All SMA	0.10 (535/122/5)	H	1.31	1.05	1.62	0.01
					All RD	0.10 (560/112/5)	H	1.15	0.92	1.44	0.22
	rs75731597	A/C	3	0.08(2957/500/34)	All SM	0.10 (1754/401/18)	H	1.27	1.10	1.47	0.001
					All CM	0.10 (975/222/9)	H	1.26	1.05	1.50	0.01
					All SMA	0.10 (520/119/7)	H	1.28	1.02	1.59	0.03
					All RD	0.09 (546/119/5)	H	1.23	0.99	1.54	0.06
FREM3	rs149914432	A/C	4	0.10(3188/685/42)	All SM	0.06 (1947/248/7)	A	0.60	0.51	0.69	8.57×10^{-13}
					All CM	0.05 (1077/141/3)	A	0.60	0.50	0.73	2.12×10^{-8}
					All SMA	0.05 (587/175/1)	A	0.58	0.45	0.74	2.26×10^{-6}
					All RD	0.05 (607/71/4)	A	0.55	0.43	0.71	9.65×10^{-7}
	rs186790584	A/T	4	0.10(3220/682/39)	All SM	0.06 (1969/241/7)	A	0.59	0.50	0.68	2.71×10^{-13}
					All CM	0.06 (1088/138/3)	A	0.60	0.50	0.72	1.44×10^{-8}
					All SMA	0.06 (593/73/1)	A	0.57	0.44	0.73	1.55×10^{-6}
					All RD	0.06 (611/69/4)	A	0.54	0.42	0.70	6.55×10^{-7}
	rs186873296	A/G	4	0.10(3216/670/40)	All SM	0.06 (1967/233/6)	A	0.57	0.49	0.66	4.23×10^{-14}
					All CM	0.06 (1087/136/2)	A	0.59	0.49	0.71	7.16×10^{-9}
					All SMA	0.05 (593/69/1)	A	0.55	0.43	0.70	4.56×10^{-7}
					All RD	0.05 (611/68/3)	A	0.53	0.41	0.69	4.67×10^{-7}
INPP4B	rs13103597	C/T	4	0.27(2120/1524/281)	All SM	0.23 (1323/773/119)	A	0.82	0.75	0.90	7.57×10^{-6}
					All CM	0.23 (733/434/62)	A	0.81	0.73	0.91	0.0001

					All SMA	0.22 (405/224/137)	A	0.76	0.64	0.90	0.001
					All RD	0.22 (409/234/40)	A	0.78	0.66	0.92	0.003
	rs77389579	G/T	4	0.06(3407/516/14)	All SM	0.04 (2019/188/2)	A	0.61	0.51	0.72	4.48×10^{-9}
					All CM	0.04 (1113/109/1)	A	0.64	0.51	0.79	1.48×10^{-5}
					All SMA	0.04 (613/52/1)	A	0.57	0.42	0.76	3.01×10^{-5}
					All RD	0.04 (625/58/0)	A	0.58	0.44	0.78	3.01×10^{-5}
ABO	rs8176719	I/D	9	0.26(2126/1506/256)	All SM	0.31 (1045/987/198)	A	0.66	0.54	0.81	8.53×10^{-5}
					All CM	0.30 (591/532/101)	A	0.64	0.50	0.83	0.0003
					All SMA	0.32 (305/322/59)	A	0.62	0.46	0.86	2.97×10^{-6}
					All RD	0.31 (314/313/56)	A	0.66	0.47	0.93	5.34×10^{-5}
	rs8176746	C/A	9	0.13(2970/878/76)	All SM	0.17 (1529/622/76)	A	1.35	1.22	1.50	9.72×10^{-9}
					All CM	0.16 (870/317/36)	A	1.20	1.06	1.37	0.004
					All SMA	0.18 (459/196/27)	A	1.43	1.23	1.67	5.83×10^{-6}
					All RD	0.18 (474/178/31)	A	1.36	1.17	1.58	0.0001
HBB	rs334	A/T	11	0.08(3320/596/33)	All SM	0.01 (2159/57/11)	H	0.16	0.12	0.21	6.81×10^{-59}
					All CM	0.01 (1185/29/8)	H	0.14	0.10	0.21	8.64×10^{-39}
					All SMA	0.01 (666/9/8)	H	0.08	0.04	0.15	2.11×10^{-31}
					All RD	0.02 (664/13/7)	H	0.11	0.06	0.20	8.41×10^{-28}
RPS6KL1	rs3742785	C/A	14	0.29(1935/1422/357)	All SM	0.30 (1032/896/187)	H	1.20	1.08	1.34	0.001
					All CM	0.29 (584/491/95)	H	1.19	1.03	1.36	0.01
					All SMA	0.32 (293/292/65)	H	1.34	1.13	1.59	0.0006
					All RD	0.28 (341/260/55)	H	0.85	0.63	1.15	0.28
HBA	α^{+}thalassaemia	A/B	16	0.40(1356/1953/637)	All SM	0.36 (853/1026/264)	A	0.82	0.76	0.89	6.59×10^{-7}
					All CM	0.37 (458/590/141)	A	0.83	0.75	0.92	0.0001
					All SMA	0.34 (275/299/65)	A	0.73	0.64	0.83	7.31×10^{-7}
					All RD	0.35 (272/311/79)	A	0.79	0.70	0.89	0.0001

† Reference/derived alleles; Chr= chromosome; MO=Model; A=additive; D=Dominant; H=heterozygote advantage; R=Recessive; OR =odds ratio; LCI and UCI= 95% Confidence interval (Lower, upper); P = P-value; All SM= All severe malaria; ALL CM=All cerebral malaria; All SMA=All severe malarial anaemia; ALL RD= All respiratory distress. Gene names highlighted in blue are related to the RBC metabolism.

Table 3.3. X- chromosome SNPs with significant associations.

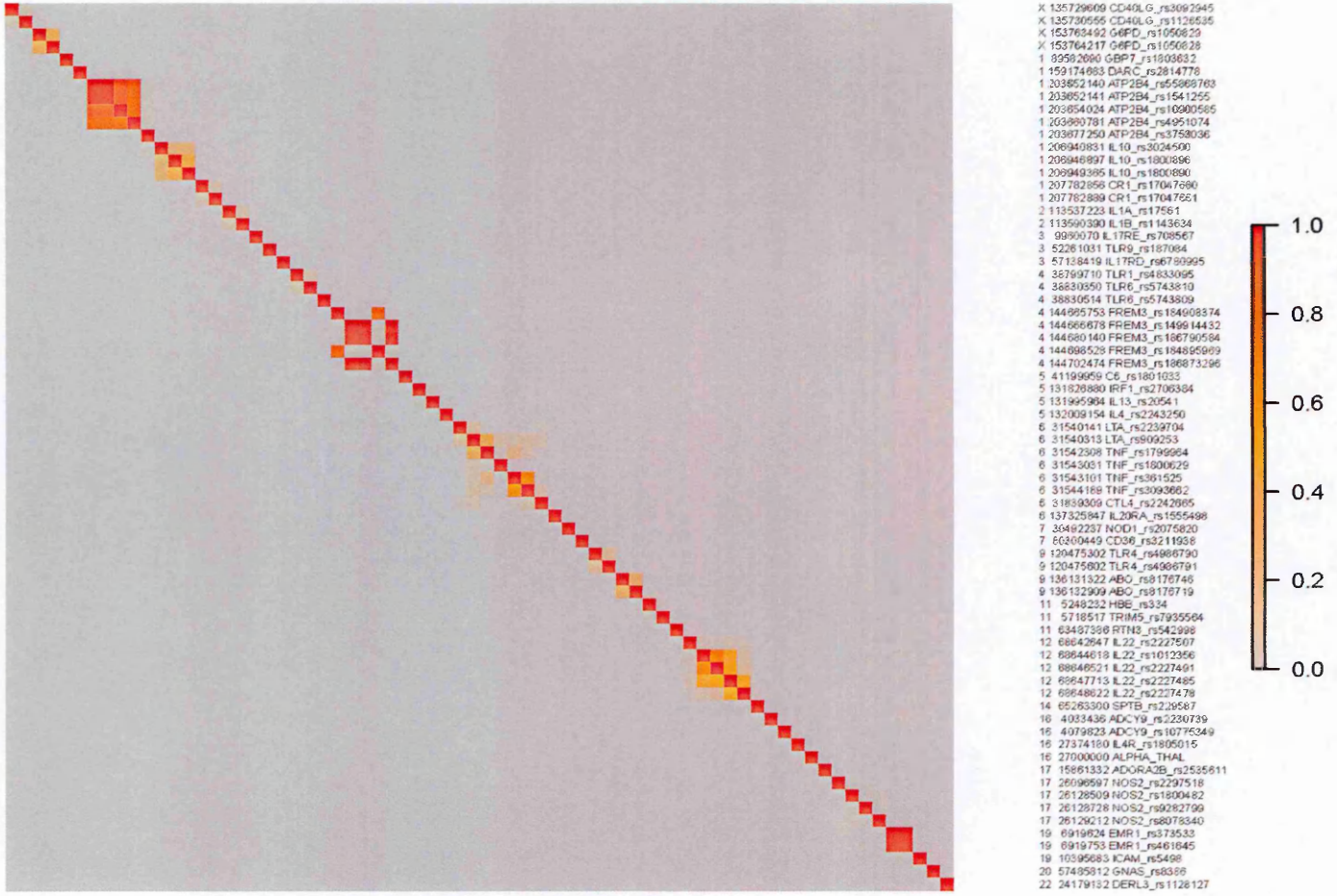
Gene	SNPID	Alleles†	Frequency of Derived allele in Controls (Norm/Het/Hom)	Sample	phenotype	Frequency of Derived allele in Cases (Norm/Het/Hom)	MO	OR	LCI	UCI	P
CD40LG	rs3092945	T/C	0.21 (2760/646/521)	ALL	SM	0.26 (1459/388/379)	A	1.14	1.06	1.22	0.0002
				ALL	CM	0.25 (815/199/209)	A	1.32	1.10	1.59	0.04
				ALL	SMA	0.26 (442/121/120)	A	1.17	1.05	1.31	0.0001
				ALL	RD	0.25 (454/128/98)	A	1.28	1.01	1.63	0.04
			0.21 (1215/646/81)	F	SM	0.25 (614/388/80)	A	1.25	1.10	1.42	0.0004
				F	CM	0.26 (355/199/48)	A	1.89	1.30	2.76	0.001
				F	SMA	0.25 (181/121/26)	A	1.33	1.10	1.62	0.003
				F	RD	0.24 (182/128/15)	A	1.29	1.01	1.63	0.03
			0.78 (440/0/1545)	M	SM	0.73 (299/0/845)	HM	0.83	0.70	0.99	0.03
				M	CM	0.76 (161/0/460)	HM	0.84	0.68	1.04	0.10
				M	SMA	0.73 (94/0/261)	HM	0.81	0.63	1.05	0.12
				M	RD	0.75 (83/0/272)	HM	0.95	0.72	1.23	0.70
G6PD	rs1050828	C/T	0.19 (2863/639/438)	ALL	SM	0.19 (1643/306/271)	H	0.82	0.70	0.97	0.01
				ALL	CM	0.18 (921/168/131)	H	0.80	0.65	0.98	0.02
				ALL	SMA	0.23 (479/90/111)	R	1.60	1.26	2.03	0.0001
				ALL	RD	0.18 (510/98/72)	H	0.90	0.74	1.09	0.28
			0.19 (1250/639/62)	F	SM	0.18 (729/306/39)	H	0.82	0.70	0.97	0.01
				F	CM	0.18 (410/168/122)	H	0.80	0.65	0.98	0.03
				F	SMA	0.17 (224/90/11)	H	0.79	0.61	1.03	0.07
				F	RD	0.19 (215/198/12)	H	0.89	0.58	1.38	0.61
			0.18 (1613/0/376)	M	SM	0.20 (914/0/232)	HM	1.11	0.92	1.33	0.26
				M	CM	0.18 (511/0/109)	HM	0.92	0.72	1.17	0.48
				M	SMA	0.28 (255/0/100)	HM	1.72	1.32	2.23	6.98×10^{-5}
				M	RD	0.17 (295/0/60)	HM	0.89	0.66	1.21	0.89

F= Female; M= Male; Models. Other abbreviations see footnote. Table 3.2.

3.4.5. Contribution of malaria candidate genes

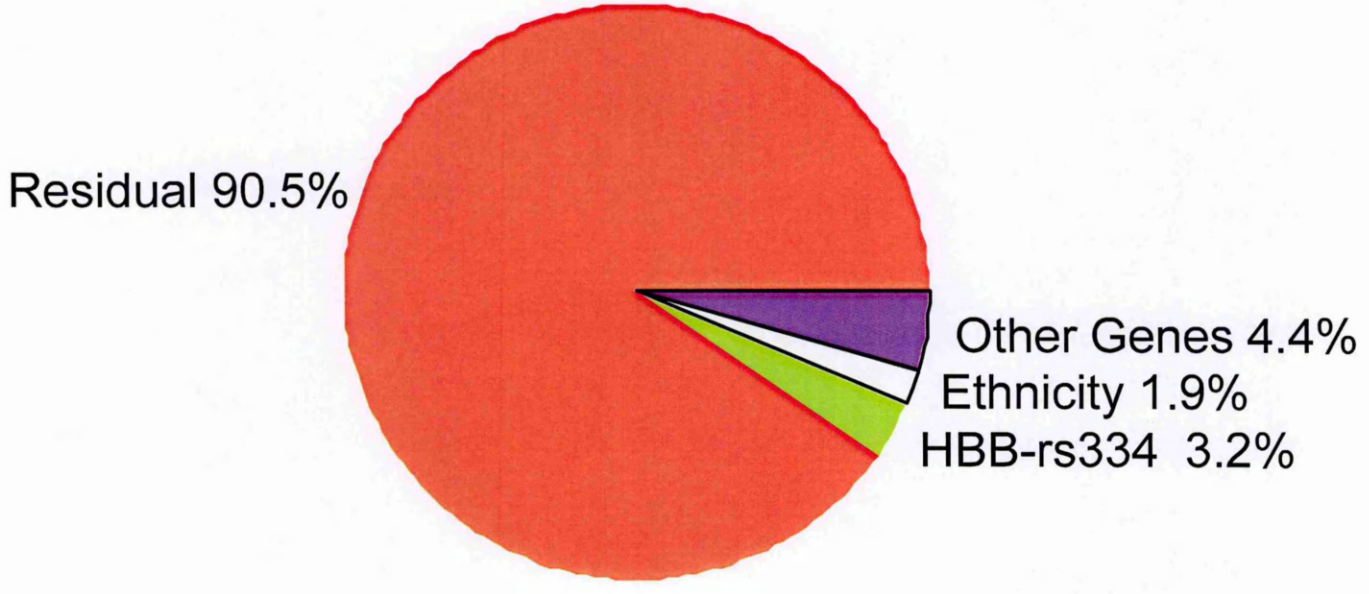
Before, calculating the genetic contribution of the variation in malaria risk it was important to examine the LD patterns of different SNPs within the same genes. This was done using a standard LD metric measure r^2 (Chapter, section 2.3.1.1). Figure 3.9 shows the LD pattern among 121 markers that were analysed. Inspecting by eye the SNPs within *ATP2B4* or the *FREM3* gene regions were in high LD. For this reason, only one SNP (the most significant) within a gene was included in the regression to estimate the variance explained. Figure 3.10 shows the proportion of variance explained by all the candidate malaria genes based on the R^2. The HbS locus contributed to the most of the total genetic variation. The total variability in the risk of SM that could be explained by all the above malaria candidate genes additively was 7.6%.

Figure 3.9. Linkage Disequilibrium among 121 loci in 71 malaria candidate gene including α^{+}thalassaemia.



The x-axis is equivalent to the y-axis and I have omitted it make the figure more visible. The colour intensity represents the strength of LD. Labels on the right indicate chromosome number, chromosome position, gene name and SNP name. Loci are ordered by chromosome number.

Figure 3.10. A Pie- Chart on variance explained by individual loci as proportion of total candidate genes.



Sixty-three individual loci across 63 candidate genes were included in this analysis.

3.4 Discussion

Genetic association studies utilising polymorphic markers in candidate genes have been successful in identifying a number of genes that are associated with susceptibility to SM. However, it must be acknowledged that the current literature contains inconsistencies and conflicting lines of evidence. With this in mind, I carried out a large case-control study of SM within a single, well characterised population, on the coast of Kenya. A major aim of this study was to reappraise previously published reports of malaria candidate-gene associations and in addition other SNPs that had potential associations in a GWAS undertaken by the MalariaGEN consortium [213]. I investigated genetic associations with SM in general and with three major sub-phenotypes of SM: CM, SMA and RD.

To minimise errors I standardised procedures using pre-defined definitions of SM and clinical sub-phenotypes, and ensured all samples underwent genotyping on the same Sequenom Mass Array platform with resulting low rates of missing data (<5%). In addition, all analyses were adjusted gender, ethnic group and the HbS polymorphism.

Based on data from my case-control study, most children presenting with severe *falciparum* malaria were less than 5 years old (88% of 2245 cases). Similar observations have been made in another group of children hospitalised for

malaria in Gabon [215]. Elsewhere, SM tends to occur in older children [216]. Differences in the age of presentation of SM may be the result of lower background immunity or other undefined factors [217]. Cerebral malaria was the most frequent feature of severity. Children with SMA had a lower fatality than children with other complications of severe malaria, these results are consistent with other studies that used hospital admission data [218, 219]. The overall case fatality rate of SM in the study was 12% (263 deaths/2245 cases), which is in keeping with other studies, where case fatality rates ranged between 8 and 40% [22, 218, 220, 221].

Notwithstanding the ascertainment bias, data from this study showed that polymorphisms affecting various aspects of RBCs (including *HBB*, *HBA*, *G6PD*, *FREM3*, *INPP4B*, *ATP2B4* and *ABO*) were among those associated with the strongest signals of differential susceptibility to SM. As expected given previous observations of the strong genetic effect of the HbS locus on malaria [46, 222], this study, confirms the known (~90%) reduced SM risk from the HbAS genotype [73] although the protective effect that I observed was smaller than in some previous reports. These differences could be due to the fact that my cases were children of young age (mean age 27 months) and thus they lacked immune basis for HbAS protection [223]. Nevertheless, despite this rigorous epidemiological evidence, the exact mechanism by which HbAS protects against

malaria is not fully understood. A recent study in Kilifi has shown that the protection afforded by HbAS increases with age, an observation consistent with enhanced acquisition of natural immunity to malaria due to the accelerated acquisition of antibodies against altered host antigens expressed on the surface of the infected RBCs [223].

The study further confirmed the known protective effects of the blood group O on life threatening malaria, which is thought to act on malaria pathogenesis through the mechanism of reduced *P. falciparum* rosetting [67, 224]. In a case-control study conducted in Mali, children with blood group O were found to be protected against SM through the mechanism of reduced *P. falciparum* rosetting [67]. Similar observations have been found in Kilifi in Kenyan children with SM [224].

The study also confirmed protective effect of α^{+}thalassaemia against severe *P. falciparum* disease. Previous studies have suggested that some of the traditional candidates might be associated with specific effects against particular sub-phenotypes of malaria. In this study however, there was little to suggest that this is true for most of the candidates, including α^{+}thalassaemia which others [41, 62, 66], have previously suggested might be particularly protective against anaemia. An explanation for this discrepancy probably relates to the sample size of these previous studies have been relatively small.

Much controversy has surrounded the question of the protective role of G6PD deficiency in malaria, and several studies have come up with conflicting results. Some studies have observed that individuals who are deficient (male hemizygotes and female homozygotes) are protected from malaria [225-227]. However, other studies have suggested that only females are protected [228]. This controversy is partly due to the fact that this disorder has three common X-linked alleles in Africa, which makes analysis and determining the phenotype difficult. In Kilifi, my colleagues have already proved that *G6PD*+202T allele is the only significant cause of G6PD [229] making results from this study more robust than a number of previous studies that have been undertaken in populations in which G6PD deficiency has subsequently been found to have a more pleomorphic aetiology. This study provides evidence that the *G6PD*+202T allele provide reduced risk for SM and SMA risk in deficient heterozygous females while increase risk to SMA in deficient hemizygous males. More recent research by the MalariaGEN consortium has found similar results [45]

In addition, to the traditional candidates, in this study, I have confirmed the associations of *ATP2B4* [230] gene, which is a membrane calcium transport protein, *FREM3* gene, which is an extracellular matrix protein and may play a role in parasite invasion and *INPP4B*, which is a protein binding gene. Both

FREM3 and *INPP4B* are immediate neighbours to the Glycophorin region (*GYPA*, *GYPB*, *GYPE)* and more likely markers for that region [213].

Besides the RBC polymorphisms, the evidence supporting other associations, mainly in cytokine or cytokine related gene polymorphisms are not far-fetched. Over the last few years, a body of literature has emerged that contend that the effects of cytokines combine to make *falciparum* malaria primarily an inflammatory cytokines driven disease [231, 232]. An excessive cytokine-mediated inflammatory response could result to SM [233]. In this present study, I did not find any association of cytokines polymorphisms with SM, given the statistical power of my study. To date, the replication of cytokine gene associations with SM has proved difficult perhaps because they are not real or because their effect sizes are small. Malaria is a highly heterogeneous disease between countries dependent on many variables such as geography, mosquitoes, rainfall and the immune status of people living in such areas all of which can be difficult to accurately quantify and account for. There is also the genetic variation of the malaria parasites between regions, although this is becoming a more tractable problem with the current genomic studies being undertaken. There is also the possibility that regional associations of cytokine polymorphism could be real as there are examples for other genes where this occurs, for example, HbC is localised to parts of West Africa, HbS to Africa and part of the Indian sub-

continent and the distribution of the different α+thalassaemia polymorphisms between Africa and Southeast Asia. Notwithstanding these factors, sample size may be one of the biggest factors and to date many studies have been of modest size. The present study is one of the biggest single-site genetic studies of severe malaria and for many of the published cytokine associations I did not find any supportive evidence for their association.

The most obvious limitation of the study was some children had more than one diagnoses on clinical examination, though these numbers were too small to affect the conclusion of our study. This problem is also shared by other similar studies on SM [218, 234, 235]. Nevertheless, this study provides useful data; to support the epidemiological and immunological studies conducted in Kilifi. Ultimately the study has also expanded our understanding of predisposing factors of SM in Kilifi and important SNPs identified may contribute to the design of future studies in the area.

In conclusion, I have conducted a large case-control study using a candidate gene approach to evaluate many previous reported associations including SNPs investigated from the MalariaGEN GWAS. More interestingly, the top putative associations with differential susceptibility to SM were to be linked to the RBC polymorphisms (*HBB, HBA, G6PD, FREM3* (surrogate for *GYP*), *INPP4B, ATP2B4, ABO*), indicative of parasite driving RBC changes. Future functional

studies of this sort should investigate the mechanism of protection afforded by these RBC polymorphisms which may provide new insights into the biology of SM that could inform the development of new approaches to prevention and treatment. Putting the above observations together raises the question of whether the above RBC- related polymorphisms are acting independently, or whether some of them are interacting to bring about protection. This will be my main focus in the next Chapter.

Chapter 4

Identifying genetic epistasis in a candidate-gene case-control study

Abstract

In recent years, a number of computational and statistical difficulties for identifying SNP-SNP interactions in high dimensional data have been studied, and several data mining approaches have been proposed. However, to the best of my knowledge the relative performance of these methods to detect SNP-SNP interactions has not been thoroughly investigated using a real data set. In this study, I directly compared the performance of three algorithms to detect gene-gene interactions in a candidate gene case-control study analysed in Chapter 3. Three methods were evaluated for their ability to detect SNP-SNP interactions: PLINK, AntEpiSeeker and SNPepistasis. Methods were contrasted on the basis of which SNPs they selected including computational cost. The results of this study demonstrate how the methods performed in detecting gene-gene interactions for a real data set, and it will be interesting to see how they perform using a large-scale data such as a GWAS study.

4.1 Introduction

One of the main primary goals of human genetics is to discover the relationships between genotypes and disease status. Although single-locus approaches have fruitfully uncovered many genetic determinants of disease susceptibility, such approaches cannot adequately explain the genetic contribution to complex diseases such as malaria. An alternative explanation is that the effects of individual genes in predicting complex phenotypes are not always necessarily independent and that epistasis contributes substantially to major human diseases [236, 237]. Identifying epistatic SNP interactions is of interest as it may enhance our current understanding of disease etiology.

As discussed in Chapters 1 and 2, there are numerous statistical methods and computational algorithms that allow investigators to test for interactions between loci. It is difficult to recommend the best algorithm without careful comparisons because of the huge difference between them, including interaction definitions, null distributions [238] and even how they include potential confounders. Several attempts have been made to evaluate different methods [208, 239-244]; however, there has been little effort to compare the performance of the different methods in the presence and absence of interactions using real data. Real data are preferable under many circumstances, since contrived models may not accurately represent complex biological processes. As it was pointed out

in the introduction chapter, there are two groups of methods in detecting epistasis, classified according to their search strategy: methods based on thorough search and methods based on search which is not exhaustive also known as stochastic search and/or greedy. In this chapter I have compared the performance of a method that I have developed (SNPepistasis, an exhaustive method) with two other well-known algorithms; PLINK (which is both exhaustive and non-exhaustive) [205], and AntEpiSeeker (which is a non-exhaustive) [207] using my case-control dataset.

4.1.2 Objectives

The central aims of this chapter are:

1. To develop a novel method that searches all possible two-locus combinations for evidence of epistatic interactions.
2. To compare the results of the novel method with other existing algorithms in the literature.
3. To identify whether polymorphisms typically act independently or whether their effects are dependent on other polymorphisms in the rural population of Kilifi.

4.2 Methods

4.2.1 Sample

In this chapter I have used the same severe malaria case-control study described in Chapter 2 section 2.1.1 that was used for the candidate single-locus association study described in Chapter 3. Overall, this dataset includes a total of 2245 SM cases and 3949 healthy controls with all individuals genotyped for 114 autosomal SNPs in 71 candidate malaria genes plus α^{+}thalassaemia 3.7kb deletion. While interactions between different X-linked SNPs and between X-linked and autosomal SNPs are of potential interest going forwards, their inclusion in this exploratory work adds a number of additional challenges including issues of power (because analyses would need to be stratified by sex) and issues relating to variable rates of Lyonisation in heterozygous females. As a consequence, I have restricted my current analyses to autosomal SNPs.

4.2.2 Statistical analysis

Initial screening of all SNPs, testing for HWE (separately in cases and controls), genotyping efficiency, and single-locus tests for association were described in Chapter 3. All the experiments were conducted on a 64-bit Windows system, with Intel ® Core™ i7-4980HQ CPU with 2.80 GHz central processing unit host (CPU) using 16 GB of memory. To measure the significance of the epistatic interaction between pairs of SNPs and SM phenotypes, three computational

methods were used: PLINK, AntEpiSeeker and SNPepistasis (see Chapter 2 section 2.3.2 for detail). All analyses were conducted with adjustment for the potential confounders ethnic group and gender. The permutation test approach was used to control for multiple testing; this estimated a p-value threshold of 0.003. I also recorded the runtimes for each analysis for each of the softwares.

4.2.2.1 Reporting interaction results

Reporting significant epistatic interactions of SNPs involve two major challenges. First, since a huge number of possible combinations are tested, a large proportion of significant associations might be false positives. In this case, I validated my results using Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database http://string.embl.de/ [245]. This tool has two main advantages: (1) It combines the information from many Protein-Protein Interaction (PPI) databases and (2) it provides a score for each interaction based on the number and quality of the reported interactions between the two potentially interacting proteins. How does the tool work? Given a PPI network, a set of genes that may be related to this network in some way, and an underlying interactome (e.g. parts from a protein-protein interaction database), the tool tries to connect the submitted genes to the network, using interactions from the PPI databases, and then reduce this extended network to only the most biological plausible interactions. In addition, computationally predicted interactions are included as these may introduce edges in the network

interaction-graph, that have not yet been fully experimentally discovered. Further detailed description of the tool can be found elsewhere [245]. The second major challenge involved interpretation of the results and developing them into biologically meaningful hypotheses. For this reason, I selected two pairs of genes with multiple lines of evidence that they genuinely interact and examined them in greater depth.

4.3 Results

4.3.1 Application to the real data

4.3.1.1 Comparison of computational time

The efficiency of the three computational methods was demonstrated by comparing runtimes. The real (clock) time required to complete the epistasis analysis were 120, 130 and 132 seconds for AntEpiSeeker, PLINK, and SNPepistasis respectively. Not surprisingly, exhaustive methods (PLINK and SNPepistasis) took a longer time to enumerate all pairs of loci, though the difference with the non-exhaustive search method (AntEpiSeeker) was small.

4.3.1.2 Statistical Performance

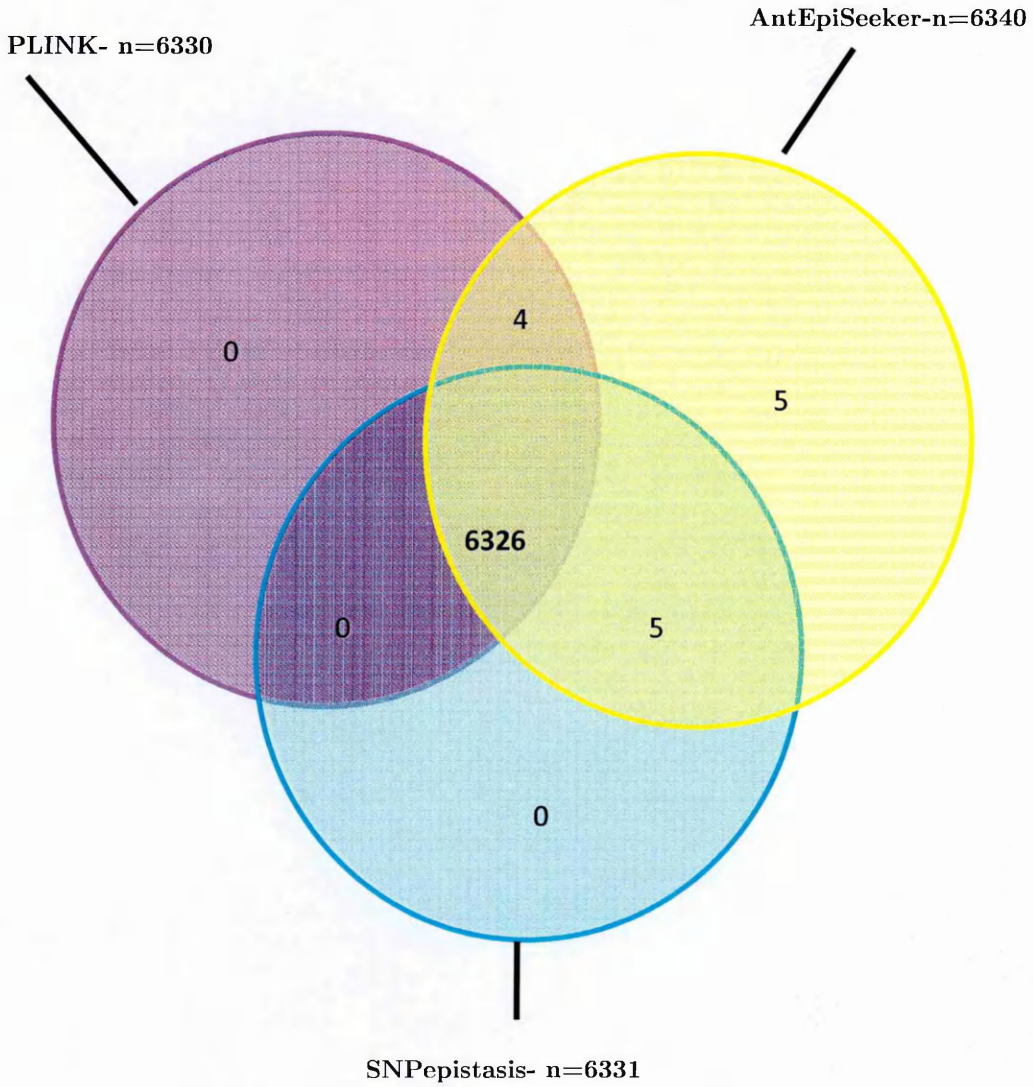
The completeness space for the three methods is shown in Figure 4.1. As there are three methods, three Venn diagrams are drawn respectively. The two methods used χ^2 test: PLINK and AntEpiSeeker while SNPepistasis used LRT to test the interaction between two SNPs. Comparing the three methods,

AntEpiSeeker identified more pair-wise interactions (6,340), than the other two methods; PLINK identified 6,330 pairs while SNPepistasis identified 6,331 pairs.

Figure 4.2 is a Q-Q plot showing that the distribution of interaction term p-values deviates from the expected uniform distribution under the null hypothesis of no epistasis (red diagonal line) for the three methods. According to the plot, the genomic inflation factors were 0.98, 1.04 and 1.02 for PLINK, AntEpiSeeker and SNPepistasis respectively, suggesting that confounding factors such as population structure have been well accounted for on all interaction analysis. As the Figure 4.2 shows, the 3 methods all produce similar plots with a low 'inflation' from the red diagonal line and I therefore assume that results are normally distributed and that the right-hand tail of deviation above the diagonal red line may harbour significant results.

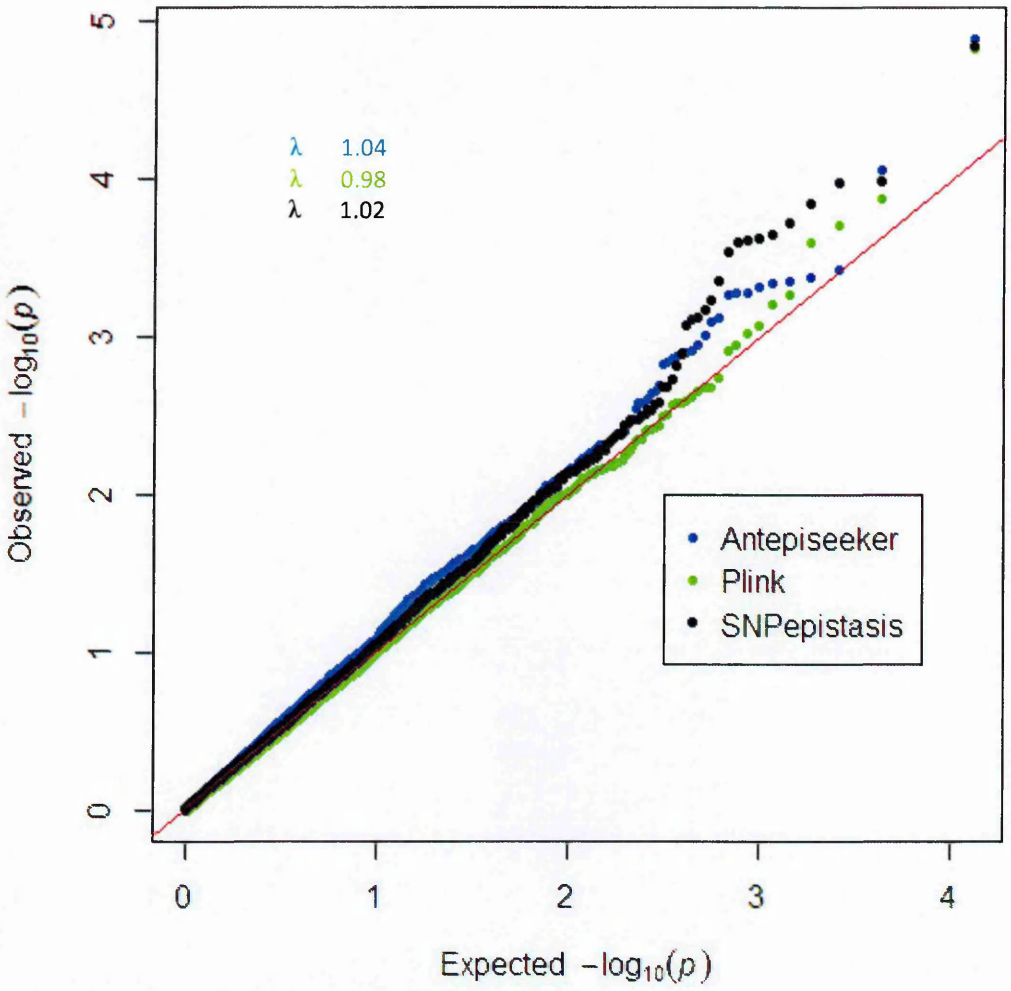
Figure 4.3 shows a Manhattan plot of the pairwise (SNPs pairs >6000) p-values derived by the three computational methods. The plot strongly suggests that there are plausible (and interesting) interactions using a threshold of $P<0.003$ (derived from permutation testing). The x-axis does not represent any particular feature or order except the SNP pair. All three methods showed similar distributions of the P-values for the SNP pairs **(Appendix D, Figure D.1)**.

Figure 4.1. The completeness interaction search space for the three computational methods.



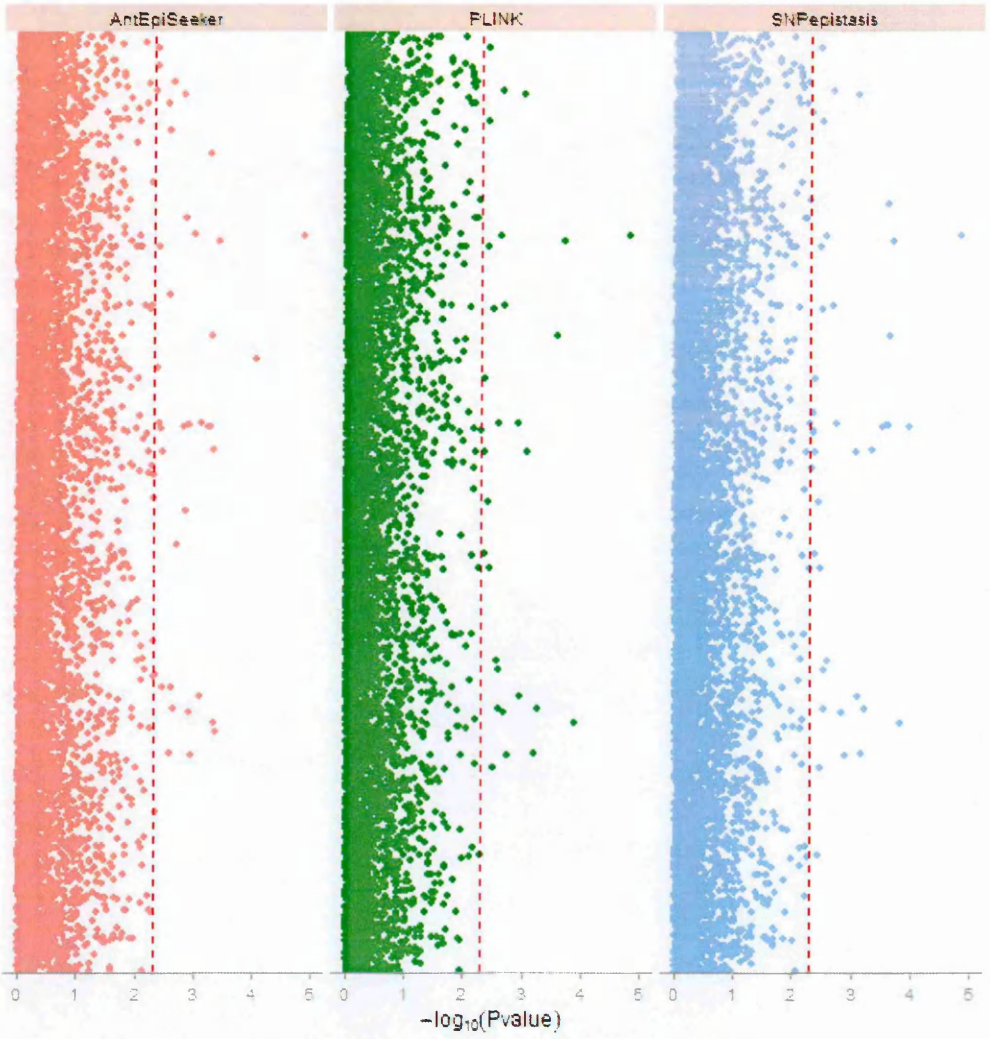
As there are three methods, three Venn diagrams are drawn respectively. PLINK and AntEpiSeeker methods used χ^2 test while SNPepistasis used LRT to test for interactions between two SNPs. PLINK identified 6,330 (6,326 + 4) pairs, AntEpiSeeker identified 6,340 (6,326+4+5+5) pairs, while SNPepistasis identified 6,331 (6,326+5) pairs.

Figure 4.2. A Q-Q plot comparing the observed $-\log_{10}$-transformed P-values from the expected uniform distribution of P-values for the three algorithms.



On the x-axis the expected uniform distribution from the χ^2 test is compared to the observed P-values from the interaction analysis derived by the three methods on the y-axis. A red line has been added to show the expected relationship between the 2 distributions under a null model of 'no-effect'. Deviation from the diagonal red line (typically y = x) can indicate aspects of data quality or some interesting interactions.

Figure 4.3. Shows a Manhattan plot of the p-values derived by the three computational methods for all 114 SNP pairs.



The dashed line represents a p-value threshold of <0.003. The X-axis depicts –log p-value and the Y-axis are the SNPs pairs.

Table 4.1 reports the top 50 SNP-pairs p values (<0.01) detected by each of the three 'interaction' methods. The Table shows for each SNP pair, their minor allele frequencies, their chromosome positions, the direction of epistatic interaction (positive or negative), represented by the interaction odds ratio (I_{OR}), and the P-values derived by the interaction. The table further shows that across the top 50 signals for each computational method the results are comparable. Furthermore the interaction identified the same top most SNP-pair interactions between *INPP4B* SNP rs13103597 located on chromosome 4 and *TNF* SNP rs1799964 located on chromosome 6 (interaction P<0.00002).

Figure 4.4 shows a graphical visualization of epistatic interactions for the top 50 hits. This method of visualization is termed as a two-way or solid edge. The nodes show the gene names and the connections show the interaction where the width of the line is proportional to the strength (Odds ratio) of the interaction. The labels in black show the ORs value. A red line signifies a negative epistatic interaction (OR >1) and a blue line depicts a positive epistatic interaction (ORs <1). For example, from the figure it can be seen that there is negative epistasis between the *HBB* gene and four other malaria candidate genes: *HBA*, *USP38*, *LPHN2* and *TNF*, and a positive epistatic interaction with the *LOC727982* gene.

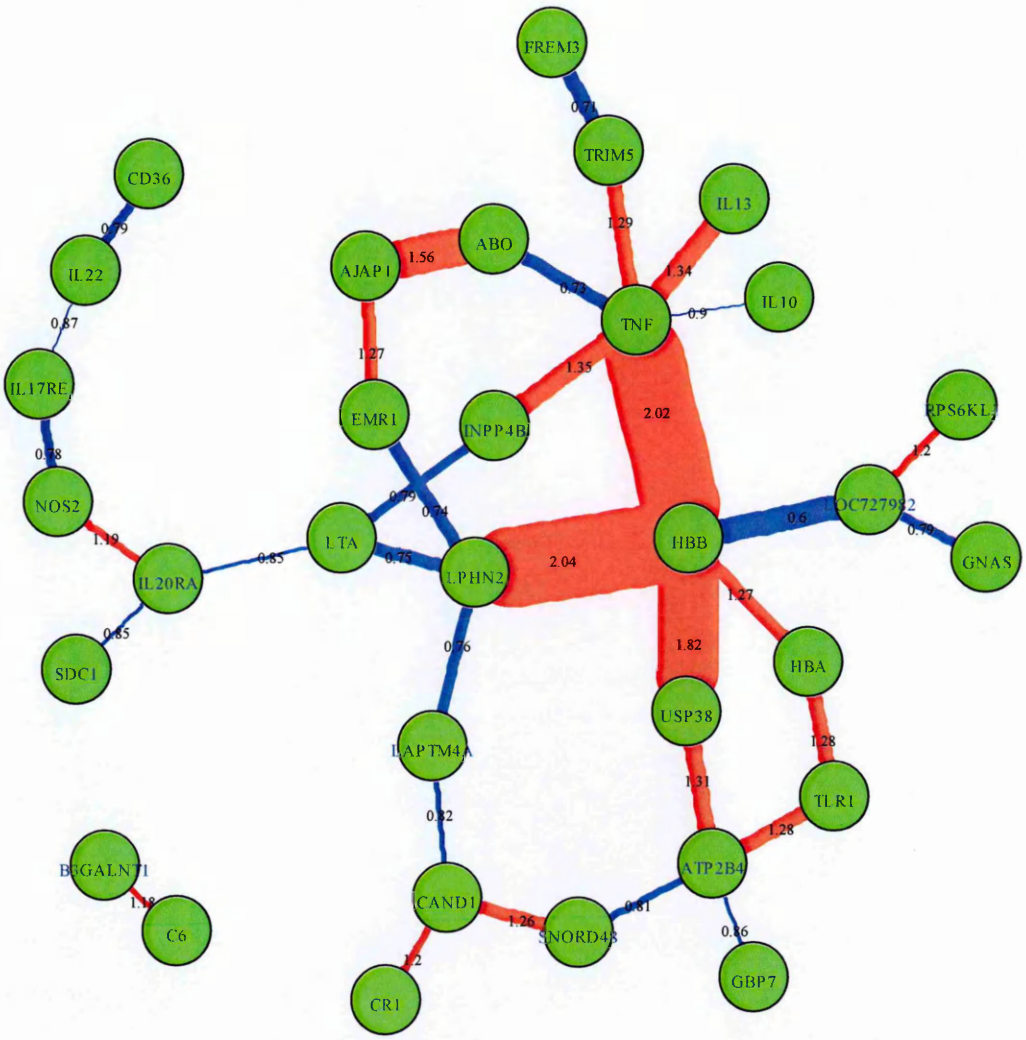
Table 4.1. Summary of the top 50 hits SNP pairs detected by PLINK, AntEpiSeeker, and SNPepistasis.

SNP 1	Gene 1	Chr 1	MAF 1	SNP 2	Gene 2	Chr 2	MAF 2	I_{OR} (95% CI)	PLINK p-value	AntEpiSeeker p-value	SNPepistasis p-value
rs13103597	*INPP4B*	4	0.25	rs1799964	*TNF*	6	0.27	1.35 (1.18-1.54)	0.00002	0.00001	0.00001
rs4951074	*ATP2B4*	1	0.32	rs4266246	*USP38*	4	0.22	1.31 (1.14-1.50)	0.0001	0.0005	0.0001
rs13103597	*INPP4B*	4	0.25	rs909253	*LTA*	6	0.49	0.79 (0.70-0.89)	0.0002	0.0004	0.0002
rs708567	*IL17RE*	3	0.50	rs8078340	*NOS2*	17	0.21	0.78 (0.69-0.89)	0.0003	0.0005	0.0002
rs55868763	*ATP2B4*	1	0.32	rs4266246	*USP38*	4	0.22	1.28 (1.11-1.46)	0.0005	0.0023	0.0006
rs1541255	*ATP2B4*	1	0.33	rs4266246	*USP38*	4	0.22	1.28 (1.11-1.46)	0.0006	0.002	0.0007
rs8386	*GNAS*	20	0.19	rs10188961	*LOC727982*	2	0.41	0.79 (0.69-0.90)	0.0009	0.003	0.0009
rs334	*HBB*	11	0.06	rs28459062	*USP38*	4	0.17	1.82 (1.27-2.62)	0.001	0.004	0.002
rs1541255	*ATP2B4*	1	0.33	rs116288147	*SNORD48*	6	0.23	0.81 (0.71-0.93)	0.002	0.001	0.001
rs55868763	*ATP2B4*	1	0.32	rs116288147	*SNORD48*	6	0.23	0.81 (0.71-0.93)	0.002	0.001	0.002
rs1555498	*IL20RA*	6	0.47	rs11899121	*SDC1*	2	0.47	0.85 (0.77-0.94)	0.002	0.005	0.002
rs13103597	*INPP4B*	4	0.25	rs361525	*TNF*	6	0.07	1.45 (1.14-1.83)	0.002	0.009	0.003
rs334	*HBB*	11	0.06	rs1800629	*TNF*	6	0.09	2.02 (1.28-3.17)	0.003	0.001	0.005
rs10459266	*CAND1*	12	0.22	rs116288147	*SNORD48*	6	0.23	1.26 (1.08-1.46)	0.003	0.007	0.003
rs12107243	*B3GALNT1*	3	0.49	rs1801033	*C6*	5	0.47	1.18 (1.06-1.32)	0.003	0.001	0.003
rs12307123	*CAND1*	12	0.23	rs973128	*LAPTM4A*	2	0.44	0.82 (0.73-0.94)	0.003	0.009	0.003
rs1555498	*IL20RA*	6	0.47	rs909253	*LTA*	6	0.49	0.85 (0.77-0.95)	0.003	0.006	0.003
rs10900585	*ATP2B4*	1	0.34	rs4266246	*USP38*	4	0.22	1.23 (1.07-1.41)	0.003	0.001	0.003
rs373533	*EMR1*	19	0.47	rs72933310	*LPHN2*	1	0.09	0.74 (0.61-0.91)	0.004	0.001	0.003
rs186790584	*FREM3*	4	0.08	rs7935564	*TRIM5*	11	0.44	0.71 (0.56-0.89)	0.004	0.004	0.004
rs8386	*GNAS*	20	0.19	rs1371478	*LOC727982*	2	0.28	0.80 (0.69-0.93)	0.005	0.006	0.004
rs461645	*EMR1*	19	0.48	rs72933310	*LPHN2*	1	0.09	0.75 (0.62-0.92)	0.005	0.004	0.004
rs6674631	*AJAP1*	1	0.12	rs461645	*EMR1*	19	0.48	1.27 (1.08-1.51)	0.005	0.003	0.004
rs334	*HBB*	11	0.06	rs10188961	*LOC727982*	2	0.41	0.60 (0.42-0.86)	0.005	0.004	0.005
rs1371474	*LOC727982*	2	0.28	rs3742785	*RPS6KL1*	14	0.29	1.20 (1.05-1.36)	0.006	0.005	0.005

rs373533	*EMR1*	19	0.47	rs72933304	*LPHN2*	1	0.08	0.74 (0.60-0.92)	0.006	0.005	0.005
rs973128	*LAPTM4A*	2	0.44	rs4650365	*LPHN2*	1	0.20	0.82 (0.71-0.95)	0.006	0.001	0.006
rs8386	*GNAS*	20	0.19	rs1371474	*LOC727982*	2	0.28	0.80 (0.69-0.94)	0.006	0.006	0.006
rs1541255	*ATP2B4*	1	0.33	rs1803632	*GBP7*	1	0.49	0.86 (0.76-0.96)	0.006	0.006	0.007
rs55868763	*ATP2B4*	1	0.32	rs1803632	*GBP7*	1	0.49	0.86 (0.76-0.96)	0.007	0.006	0.007
rs334	*HBB*	11	0.06	rs72933304	*LPHN2*	1	0.08	2.04 (1.22-3.42)	0.007	0.009	0.01
rs973128	*LAPTM4*	2	0.44	rs72933310	*LPHN2*	1	0.09	0.76 (0.63-0.93)	0.007	0.005	0.006
rs186873296	*FREM3*	4	0.08	rs7935564	*TRIM5*	11	0.44	0.72 (0.57-0.92)	0.007	0.005	0.006
rs461645	*EMR1*	19	0.48	rs72933304	*LPHN2*	1	0.08	0.75 (0.61-0.92)	0.007	0.005	0.007
rs10459266	*CAND1*	12	0.22	rs17047661	*CR1*	1	0.33	1.20 (1.05-1.38)	0.008	0.008	0.008
rs1555498	*IL20RA*	6	0.47	rs8078340	*NOS2*	17	0.21	1.19 (1.05-1.35)	0.008	0.005	0.008
rs4650365	*LPHN2*	1	0.20	rs2239704	*LTA*	6	0.49	0.75 (0.60-0.93)	0.008	0.005	0.007
rs1566830	*CAND1*	12	0.26	rs973128	*LAPTM4A*	2	0.13	0.85 (0.75-0.96)	0.009	0.005	0.009
rs1800629	*TNF*	6	0.09	rs7935564	*TRIM5*	11	0.44	1.29 (1.06-1.57)	0.01	0.02	0.01
rs1371478	*LOC727982*	2	0.28	rs374278	*RPS6KL1*	14	0.29	1.19 (1.04-1.35)	0.01	0.02	0.01
rs708567	*IL17RE*	3	0.50	rs2227491	*IL22*	12	0.41	0.87 (0.78-0.97)	0.01	0.03	0.01
α^+Thal	*HBA*	16	0.39	rs334	*HBB*	11	0.06	1.27 (1.06-1.53)	0.01	0.02	0.01
rs20541	*IL13*	5	0.22	rs1800629	*TNF*	6	0.09	1.34 (1.07-1.70)	0.01	0.03	0.01
rs8176746	*ABO*	9	0.15	rs3093662	*TNF*	6	0.11	0.73 (0.58-0.93)	0.01	0.03	0.01
rs1541255	*ATP2B4*	1	0.33	rs4833095	*TLR1*	4	0.09	1.29 (1.06-1.59)	0.01	0.02	0.01
α^+Thal	*HBA*	16	0.39	rs4833095	*TLR1*	4	0.09	1.28 (1.06-1.57)	0.01	0.02	0.01
rs149914432	*FREM3*	4	0.08	rs7935564	*TRIM5*	11	0.44	0.74 (0.59-0.94)	0.01	0.04	0.01
rs3211938	*CD36*	7	0.10	rs1012356	*IL22*	12	0.49	0.79 (0.66-0.95)	0.01	0.01	0.01
rs8176719	*ABO*	9	0.28	rs6674631	*AJAP1*	1	0.08	1.56 (1.10-2.20)	0.01	0.02	0.01
rs3024500	*IL10*	1	0.46	rs1799964	*TNF*	6	0.27	0.90 (0.76-0.97)	0.01	0.02	0.01

The rows of the table are sorted in ascending order of interaction p-values for PLINK (column 10). SNP= single nucleotide polymorphism, Chr=Chromosome, MAF =minor allele frequency, I_{OR}= Interaction Odds ratio.

Figure 4.4. A two-way epistatic interactions network among the top 50 hits.



Nodes in green colour are the gene names. Labels in black show the strengths of interaction effects and the edge width is proportionally adjusted to the ORs value. A red line signifies a negative epistatic interaction (OR > 1) and a blue line depicts a positive epistatic interaction (ORs <1). The data for PLINK are shown in this figure.

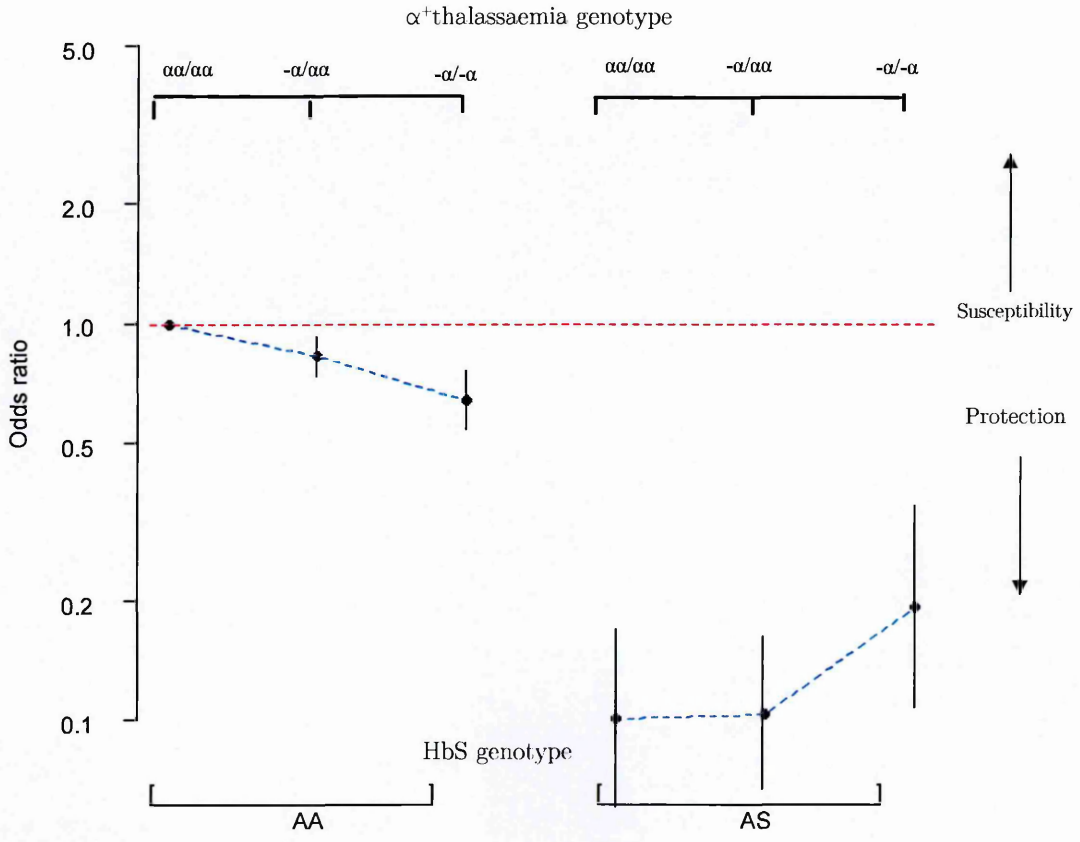
4.3.2 Biological Interpretation

One of the biggest challenges in examining epistatic interactions between genes or SNPs is the interpretation of the results. To begin, I have asked myself three basic conceptual questions:

1. What do all these statistically significant interactions mean?
2. Are they all real and do I believe them?
3. Which of the many significant gene interactions are useful for predicting who is at risk for SM?

There are all basic questions that require critical thinking. Nevertheless, I still need to interpret these results in the context of human biology so that they may be used for the benefit of disease treatment and prevention. Here, I focused on two different approaches. First, as a proof of principle, I selected two pairs of variants (HbS and α^{+}thalassaemia) with multiple lines of biological evidence that they genuinely interact and examined them in greater depth [57, 62, 103]. Returning back to the epistatic interaction results detected by the three computational methods (Table 4.3 and Figure 4.4), they all confirmed (AntEpiSeeker P=0.02, PLINK, P=0.01 and SNPepistasis P=0.01) some evidence of epistatic interactions between HbAS and α^{+}thalassaemia with SM. To aid interpret these results, I used a Forrest plot to illustrate the effect of the interaction between HbS and α^{+}thalassaemia compared (Figure 4.5).

Figure 4.5. A Forrest plot for epistasis between HbAS and α^{+}thalassaemia for association with SM.



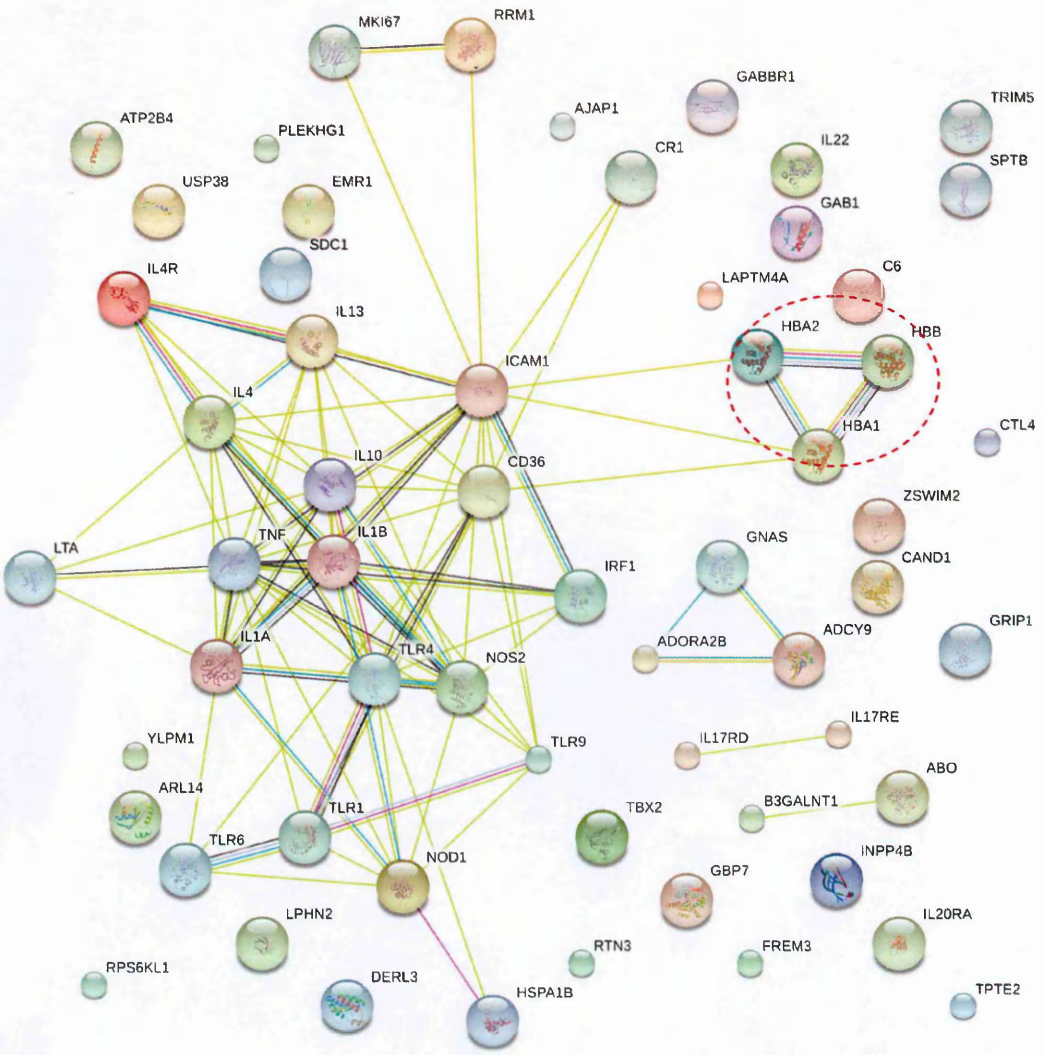
Allele group	a	b	c	d	e	f
Cases	828	1002	252	16	22	13
Controls	1131	1649	537	216	288	92
P-value	ref	0.002	5.8×10^{-7}	3.1×10^{-18}	1.4×10^{-23}	4.0×10^{-8}

For each combination of genotypes (x-axis), I computed the odds ratio ±95% CI (y-axis) and p-value relative to the reference group (α^{+}thalassaemia: **αα/αα**/HbAA). The red line shows the point of no effect. The odds ratio has been plotted on a log scale so that the distances above and below 1 represent the same size of effect although opposite effect. I also list the sample size for cases and controls. Three methods have been used to detect this epistasis: AntEpiSeeker (P=0.02), PLINK (P=0.01) and SNPepistasis (P=0.01). The data for SNPepistasis are shown in this figure. The protection afforded by both polymorphisms when inherited together is reduced.

Allele group "a" shows the baseline for normal α^+thalassaemia and HbAA. As shown in allele groups "a", "b" and "c", on a normal HbAA background, α^+thalassaemia tends to protects reaching an OR of 0.6 in the homozygote state. However, this trend for protection is not seen in individuals with HbAS, indeed the α^+thalassaemia locus reduces the effect of HbAS (OR 0.1 to 0.2).

The second approach was to check if epistatic interactions reported from my analysis had evidence from the literature and biological databases. The tool used for this was the STRING database. I submitted a list of protein names included in my study to the database for construction of a protein interaction network as seen in Figure 4.6. Interestingly, the STRING tool demonstrated that haemoglobin related proteins: *HBA1*, *HBA2* and *HBB* for which there was evidence of epistatic interaction clustered together and formed an interaction network. I further compared the top 50 interactions as reported in Figure 4.4 with the nodes of the interaction network generated by the STRING program (Figure 4.6). From this comparison, 3 interactions gene-pairs from Figure 4.4 matched interacting nodes from the STRING database. These gene interaction pairs were: *TNF-IL10*, *TNF-IL13* and *IL10-IL13*. Moreover, this seems like a higher order (3-way) interaction. The presence of these interactions in STRING database appears to validate my results.

Figure 4.6. Illustrates the protein-protein interaction network generated from STRING.



The graph was generated using STRING. Edges in this graph only mean that in some publication evidence was found that the connected genes interact. There is no temporal or directional component given in this graph. Purple lines indicate evidence from experiments, blue lines indicate homology of sequences, green lines indicates evidence from text-mining, dark green node indicates association by neighbourhood, Red node indicates association by gene fusion whereas a blue node indicates association in gene and protein databases. Haemoglobin related proteins: *HBA1*, *HBA2* and *HBB* for which there is evidence of epistatic interaction is marked with a red circle.

4.4 Discussion

It is now widely accepted that multiple genes may be responsible for many complex diseases and as such in the study of such diseases, emphasis is now been placed on finding all possible epistatic interactions between these genetic variants [246, 247]. By allowing for epistatic interactions between potential disease loci, we may succeed in identifying genetic variants that might otherwise have remained undetected. In addition, the identification and characterization of epistatic interactions might provide new knowledge on complex genetic mechanisms that underlie complex diseases [248] such as malarial disease.

Methods of analysis that detect or control the phenomenon of epistasis are clearly of growing interest in the genetic dissection of human complex disease [249]. Despite advances in speed, the most common bench-mark for methods that detect epistasis remains simulated data, where single epistatic interaction embedded in a small number of SNPs is used to judge a method's power and false positive rate under various parameter settings. In this work, I evaluated three statistical methods (PLINK, AntEpiSeeker and SNPepistasis) for detecting epistasis using a real life dataset. SNPepistasis was my novel method developed using a standard logistic regression framework while the other two methods were readily available from the literature.

This study represents the first of its kind to compare the three methods using a real life dataset of SM. From a practical point of view, the main differences between these three methods were: the computational time required to complete the analysis and data manipulation before final analysis. From a methodological perspective, it is of note that PLINK, AntEpiSeeker and SNPepistasis detected the same top SNP pair despite different configurations of parameter settings. AntEpiSeeker employs an evolutionary computation non-exhaustive search strategy in comparison to the exhaustive search strategy of PLINK and SNPepistasis, it is a more satisfying outcome that the three methods produced similar results.

The top most significant interaction pair (interaction $P=1.0\times10^{-5}$) included the *INPP4B* SNP rs13103597 located on chromosome 4 and *TNF* SNP rs1799964 located on chromosome 6. The role of *INPP4B* was discussed in Chapter 3 section 3.4 while *TNF* is a cytokine that is thought to play a major role in imparting pathogenicity to malaria [250]. It is also implicated in playing a central role in CM pathogenesis by inducing changes in cerebral endothelial cells, leading to the surface expression of adhesion molecules, thereby enhancing parasitized erythrocyte sequestration. It is worth noting that in the single-SNP association analysis the *INPP4B* rs13103597 locus showed a significant protective effect (OR=0.82, $P=7.57\times10^{-6}$) against SM (Table 3.2), while no

association was observed between the *TNF* rs1799964 locus (OR=0.95, P=0.39), indicating that if I had focused on SNPs with main effect to test for interaction I would have missed this significant interaction. The same holds true for a recently published study of an exhaustive genome-wide analysis [251], whereby the study indicated that significant epistatic interactions would have been missed if SNPs that did not display any main effect had been excluded a priori [251]. However, it remains unclear whether this interaction would impact on disease progression. There is a need to explore the possible role of this interaction in malaria pathogenesis.

Previous studies have shown that an important causative protein in malaria is haemoglobin, which has two variants – heterozygote HbAS and homozygote HbSS that can cause sickle cell anaemia. While HbSS is a lethal mutation leading to premature death, individuals with HbAS are protected against malaria. Additionally, there exist other mutations that can protect against severe and fatal malaria – heterozygote $-\alpha/\alpha\alpha$ and homozygote $-\alpha/-\alpha$, which cause α^{+}thalassaemia. A negative epistasis has been reported between HbAS and α^{+}thalassaemia in a Kilifi Cohort [103] and this has been replicated in multiple studies (see Chapter 1 section 1.5.1.1). As a starting pointing in understanding the observed interactions results, I started by investigating this known epistasis interaction as a proof of principle. Interestingly, the present study confirmed

this epistasis using a different design (case-control) to that used previously (cohort) [103], although the strength of association was rather weak.

To complement my results, I used STRING database [245] , to explore the wide spectrum of all biologically plausible authentic epistasis observed in my present study. Given the many hits in Table 4.1; it was a surprising fact that very few of them overlapped with the reported interactions evidence available in STRING. This paradox may be explained by: 1) some of these interactions represent statistical but not biologically plausible interactions; 2) the STRING database may yet have to be updated with evidence of these genetic interactions; 3) they might be novel epistatic interactions yet to be discovered. Nevertheless, the STRING tool demonstrated evidence of epistatic interactions among haemoglobin related proteins (*HBA1*, *HBA2* and *HBB)* and some cytokines genes (*TNF*, *IL10*, and *IL13*) relevant to malaria immunity. This is consistent with the conclusion that the STRING database is an important tool in interpretation of statistical data and in validating hypotheses of interaction between plausible genes [252].

In summary, I have compared and contrasted the performance of three selected approaches to identify epistatic interactions between SNPs using a real dataset of SM. While I have not conducted a comprehensive comparison in order to determine which method(s) performed the best for identifying interactions, each

method has potential for application to high-dimensional genetic interaction data with any phenotype of interest. While this study has identified some interesting interactions, this remains a preliminary analysis at the beginning stage of applying epistasis using a candidate gene study. It must be emphasized here that it only included a small set of SNPs, and this could be a major limitation to drawing strong inferences about the epistasis observed if they are real. Nevertheless, they have served to illustrate the process and complexities of detecting epistasis. In Chapter 6, I will use a large-scale GWAS dataset to characterise the process of detecting epistasis interactions.

Chapter 5

The haplotype structure and patterns of linkage disequilibrium across the HbS and α^+thalassaemia 3.7kb locus in the Kilifi population

Abstract

Surprisingly little work has been done to characterise the haplotype structure and LD patterns surrounding some of the well-known malaria susceptibility polymorphisms such as HbS and α^+thalassaemia. One intriguing observation from the candidate gene case-control study (Chapters 3 and 4) was the protective and co-existence roles of these two variants in SM. This observation informed my decision to investigate comparatively their haplotype structures and their patterns of LD purposely to throw more light on how the two variants have evolved in the Kilifi population. First, I carried out a high-resolution analysis of the haplotype structure across a 1-megabase genomic region on chromosome 11 using 1964 SNPs and across 300 kilobases on chromosome 16 using 210 SNPs. Next, I evaluated the extent of LD among biallelic markers surrounding the two

variants using a correlation matrix and pair-wise LD measures (r^2 and D'). To the best of my knowledge, this is the first study to comprehensively explore the haplotype structure and LD patterns of the HbS and α^+thalassaemia variants that are found to co-exist in Kilifi. The findings described in this study may be a valuable resource for future functional studies designed to refine the genomic structure in both HbS and α^+thalassaemia genomic regions.

5.1 Introduction

Recent DNA studies of modern human chromosomes have largely used patterns of modern genetic diversity to infer population histories, by applying simple models of chromosome evolution [253]. The evolutionary history is better studied by analysing haplotype of DNA sequences. Ultimately, there is great interest in understanding the haplotype structures in the human genome using identified genetic markers because: 1) haplotype structures may be used to infer the human evolutionary history and localise genetic variants underlying complex diseases ; and 2) recent advances in the high- throughput genotyping technology now make it increasingly possible to study millions of genetic markers in population samples of reasonable sizes [254]. Since the alleles at closely linked markers on a single chromosome are sometimes in LD, one important aspect is to identify the LD patterns in a region of interest or a population. LD patterns observed in natural population are the results of a complex interplay by many

factors, including population history, recombination rates, gene conversion, natural selection, and other factors [255]. Therefore, it is not surprising that there is a much variation in LD among different genomic regions and population.

As reviewed earlier in the introduction Chapter 1, the HbS-rs334 and α^+thalassaemia are inherited disorders of haemoglobin. HbS is a variant of the *HBB* gene located on chromosome 11 while the α^+thalassaemia deletion is located on chromosome 16. These two chromosomes are highly polymorphic each with a high density of genes. Although biomedical research has accelerated its pace in the past decades, little has been done to look at how the haplotype structure and LD patterns can be utilised to aid our understanding of the dynamics of the malaria-driven selection of these two variants. Although, restriction fragment length polymorphism (RFLP) based haplotype analyses have been performed for both HbS and α^+thalassaemia [256-258] along with looking at some sequences for short regions sequence data [259], little is known about how the two variants have evolved in the Kilifi malaria-endemic area and in larger chromosomal regions around the genes. Moreover, neither HbS nor α^+thalassaemia are typed on the current GWAS chips. Having useful marker on the chips that are in LD with the causal variants could help infer/impute the HbS and α^+thalassaemia genotypes, saving the time, effort and cost of direct genotyping, especially if the assays are non-trivial.

5.1.1 Objectives

The main objective of this chapter was to study the dynamics of evolutionary selection of HbS and α^+thalassaemia gene regions. To achieve this goal, more specific objectives were set as follows:

1. Characterise the haplotype structure and sub-groupings within and between ethnic groups in the two gene regions.
2. Search for a functional SNP strongly correlated or with high LD with HbS or α^+thalassaemia deletion.

5.2 Material and methods

5.2.1 Study population

In order to capture most of the diversity in the Kilifi population, the study sample comprised of children with SM and healthy controls drawn from the Kenya GWAS study of SM (see Chapter 2, section 2.2.3).

5.2.2 Genotyping and SNP selection

Samples passing the QC were genotyped using the Illumina HumanOmni2.5-4 genotyping chip while identification of the α^+thalassaemia genotype was done as described in chapter 2, section 2.2.1.3. To generate the datasets, I selected all common SNPs (>5% minor allele frequency) spanning in a 1Mb region on chromosome 11 and a 300kb region on chromosome 16. A total of 1,964 and 210 polymorphic autosomal SNPs were included from chromosome 11 and 16 respectively. The data has been phased so there were two haplotypes available for each individual, and two alleles at each SNP, encoded as 0 (reference) and 1 (derived).

5.2.3 Statistical analysis

Genotype frequencies of the HbS and α^+thalassaemia were determined by direct allele counting. Descriptive statistics was used for basic characterisations of the study population. The haplotype structures were analysed using the R function Hclust (http://cran.r-project.org/) for all subjects (cases and controls combined). To characterise the haplotype structure, individuals were used who were identified as being homozygous either for HbS, α^+thalassaemia or having neither polymorphism. I did not use heterozygous individuals to reduce the chance of incorrect assignment of markers to haplotypes.

Patterns of haplotype homogeneity in the chromosomes that have undergone selective pressure (HbS and $-\alpha$) were compared with the patterns in the non-selected chromosomes (HbA and $\alpha\alpha$) respectively. Genotypes were constructed from the haplotype data before computing pair-wise correlation coefficient matrix (r). The tag SNPs were selected based on evidence from the correlation matrix between the SNP and HbS or α^+thalassaemia loci. Pair-wise LD was estimated by calculating pairwise D-prime (D') and R-squared (r^2) metrics using customised R scripts. Principal component analysis (PCA) was used to determine the sub-structure within the Kilifi population. Receiver operating characteristic (ROC) curve analysis was used to assess predictive accuracy of chip intensiy data in predicting α^+thalassaemia genotypes.

5.3 Results

5.3.1 General data of the studied population

A total of 2,842 (both cases and controls) children were included: mean age 18.8 ($\pm$20.2) months, 1383 (49%) were female and 1459 (51%) male. The study population comprised a variable mixture of three main ethnic groups: Chonyi 32% (n=901), Giriama 51% (n=1459), Kauma 9% (n=258), and other 8% (n=224). Children belonging to other ethnic origin were excluded from further analyses.

5.3.2 Frequencies of the HbS and α^+thalassaemia polymorphisms

Both HbS and α^+thalassaemia polymorphisms occur at high frequencies in the Kilifi malaria-endemic population as shown in Table 5.1. The overall frequencies of HbAA and HbSS were 89% and 1% respectively while those of $\alpha\alpha/\alpha\alpha$ and $-\alpha/-\alpha$ were 37% and 14% respectively. The frequency of HbSS was higher in Chonyi (2%) compared with Giriama (0.4%) and Kauma (0.8%) ethnic groups (Table 5.1) while the frequency of $-\alpha/-\alpha$ genotype was slightly higher in Giriama (16%) subjects as compared to the other two ethnic groups (Table 5.1).

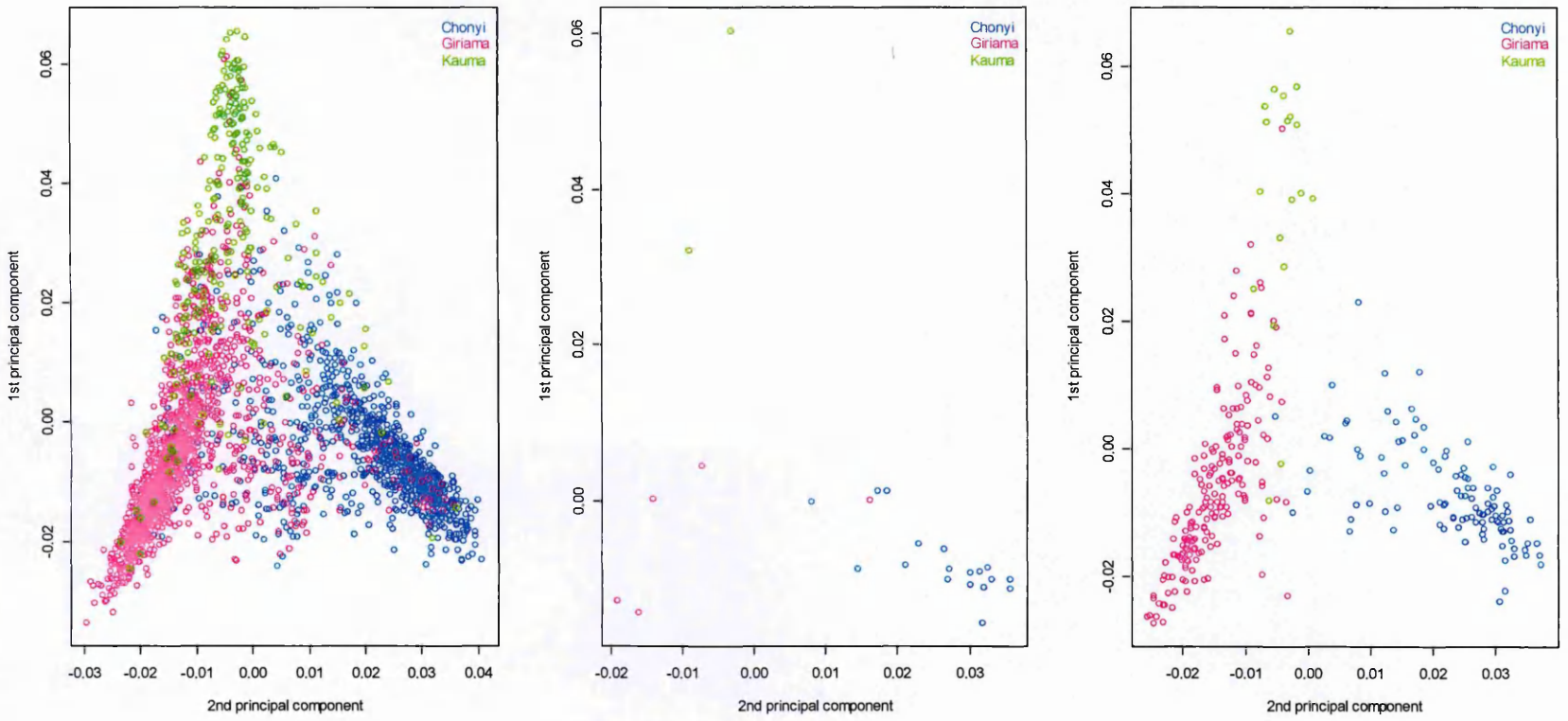
5.3.3 Inference of population substructure

To investigate the relationship between the population structure and self-reported ethnicity, I performed a PCA analysis for all 2,618 individuals, and a two-dimensional plot based on the top two principal components (PC1 and PC2) as shown in Figure 5.1a. I then coloured the plot according to ethnicity. As expected the first two (PCs), clearly separated the Chonyi, Giriama and Kauma. Nevertheless, some individuals could be confidently assigned to a specific ethnic group, whereas others seemed to have a complex ancestry. In Figures 5.1b and 5.1c, I have only shown the homozygous HbS and α^+thalassaemia individuals respectively used for the haplotype analysis. As can be seen, the majority of individuals are mostly separated from each other and located in the distinct ethnic groups; Chonyi, Giriama or Kauma.

Table 5.1. Frequencies of the HbS and α^+thalassaemia polymorphisms in the Kilifi population.

Polymorphisms	Overall N=2845	Ethnic group		
		Chonyi	Giriama	Kauma
HbS				
Homozygous wild-type (HbAA)	2535 (89%)	762 (85%)	1344 (93.4%)	220 (85.9%)
Heterozygous (HbAS)	262 (9%)	117 (13%)	90 (6.2%)	34 (13.2%)
Homozygous carriers (HbSS)	28 (1%)	18 (2%)	5 (0.4%)	2 (0.8%)
α^+thalassaemia				
Homozygous wild-type (αα/αα)	1007 (37%)	302 (35%)	586 (36%)	106 (42%)
Heterozygous (–α/αα)	1340 (49%)	456 (52%)	666 (48%)	115 (46%)
Homozygous (–α/–α)	387 (14%)	115 (13%)	218 (16%)	29 (12%)

Figure 5.1. Principal Components Analysis of individuals in the Kenya GWAS of severe malaria.



The x-axis denotes the value of PC2, while the x-axis denotes the value of PC1, with each solid circle representing one individual, and the colour is assigned according to self-reported ethnicity. The Dataset consisted of 2618 individuals and 1964 SNPs after thinning by LD. (a) All individuals; (b) individuals homozygous for HbS; (c) individuals homozygous for α^{+}thalassaemia.

5.3.4 The haplotype structure

Figure 5.2 shows the extent of homogeneity in the major haplotypes observed in HbS and HbA chromosomes in the samples analysed. The chromosomes are represented on the y-axis and 1,964 SNPs that were selected across the 1-Mb on the x-axis. For each SNP, the ancestral state is coloured black, while the derived allele is coloured white. I observed two features in the HbS chromosomes in this population, which may together be favourable for future genetics studies. First, the HbS chromosomes were extensively homogenous over the 1-Mb region though there seem to be two distinct haplotypes existing in this population (Figure 5.2). Second, examining the haplotype patterns there seem to be one 'common' haplotype (200-kb region -part marked with a green box in Figure 5.2) around the HbS locus represented by the vertical red line. Common here means represented across the majority of chromosomes. The HbA chromosomes however, diversified essentially throughout the region as shown in Figure 5.2.

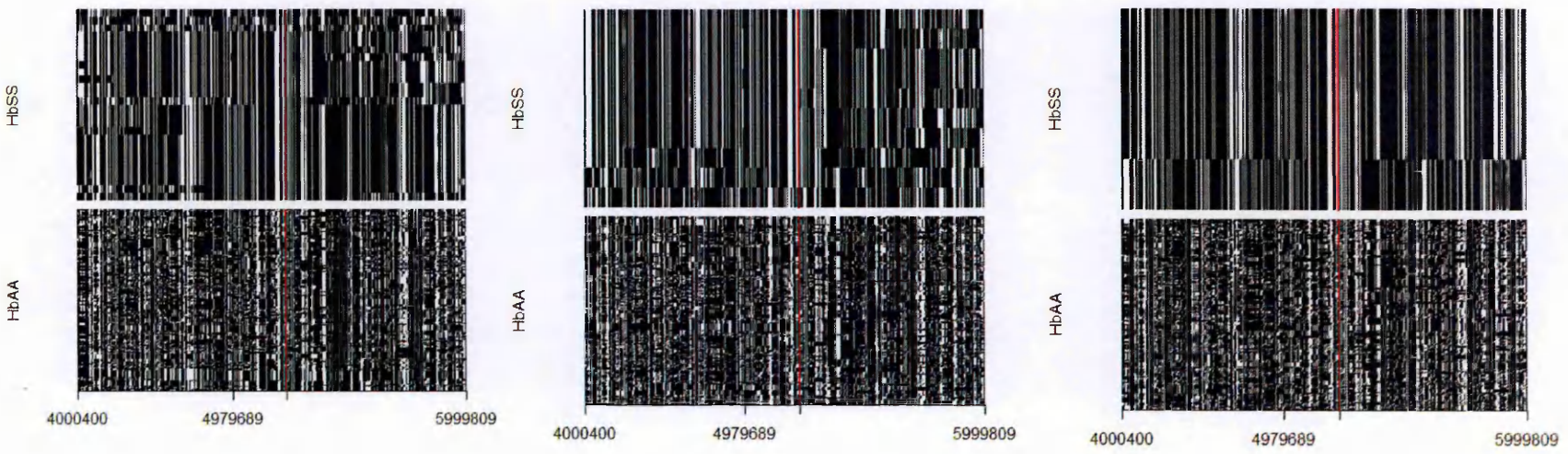
Figure 5.3 illustrates patterns of haplotype structure in the HbS and HbA chromosomes observed in the study population stratified by the three major ethnic groups.. The most common HbS haplotype maintained its structure around HbS locus (200-kb region). Two major haplotype structures were observed for the HbS chromosomes in all the ethnic groups.

Figure 5.4 shows the extent of heterogeneity in the major long-range haplotypes observed in αα and −α chromosomes. The location of the α^+thalassaemia deletion in both plots is shown by the vertical red line. Unlike *HBB*, no apparent structure was observed in either αα or −α chromosomes across the 300-kb region. Figure 5.5 illustrates patterns of haplotype structure in the αα and −α chromosomes stratified by the three major ethnic groups. The haplotype patterns were not distinct among the three ethnic groups in the αα chromosomes but some differences were observed in the −α chromosomes by ethnicity.

Figure 5.2. Illustration of the haplotype structure in HbS and HbA chromosomes observed in the study population.

Genetic distance is indicated at the bottom of the plot in kilo-bases (kb) with respect to the chromosomal location. Each row represents a chromosome from an individual and the columns represent the SNPs for a 1-Mb region across the HbS locus (represented by the red vertical line). The gene related to the HbS locus is identified by the red box in the lower panel. For each SNP, the major allele is in black and minor allele in white. The SNPs are arranged according to the position in chromosome 11 region. In total there were 28 individuals contributing to the HbS chromosomes. In order to maintain clarity for the HbA chromosomes, 400 individuals were selected at random. The lower panel of the plot shows genes in the region. Gene names follow the HUGO classification. Gene structure is shown by the blue boxes (exons) and the blue lines (introns). The gene transcription direction is shown by the arrows. All gene positions/locations are referenced to GRCh37.

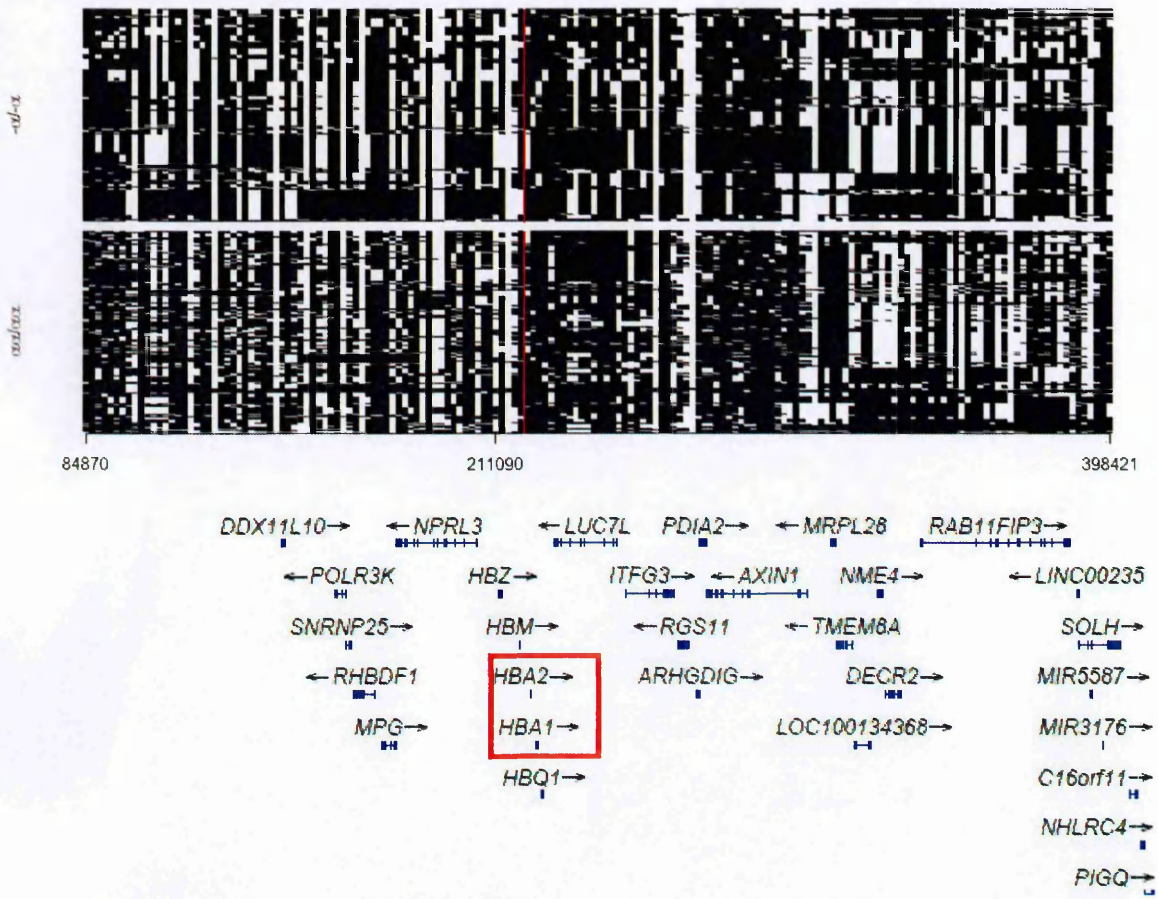
Figure 5.3. Illustration of the extent of haplotype homogeneity in the HbS and HbA chromosomes observed in the study population stratified by ethnicity.



(a) Chonyi **(b) Giriama** **(c) Kauma**

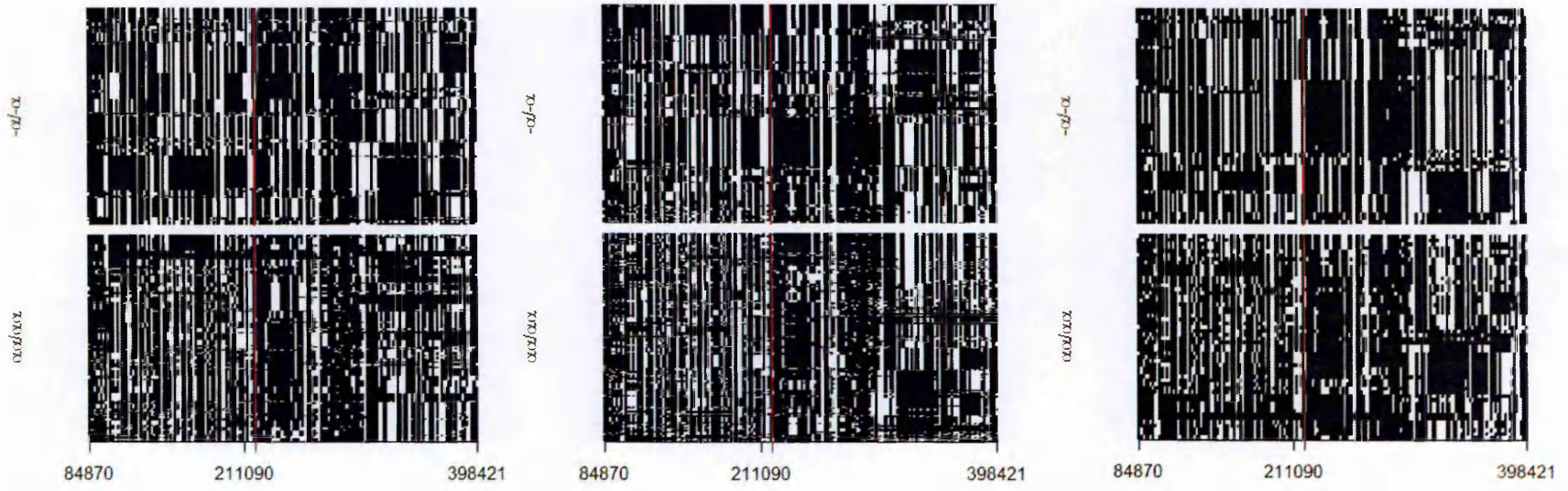
It is worth noting that unequal sample sizes of the HbS and HbA chromosomes were used here. Figures (a), (b) and (c) show the haplotype homogeneity in the HbS and HbA chromosomes observed in the Chonyi, Giriama and Kauma ethnic groups respectively. For further description of the axis and the data, see footnote, Figure 5.2.

Figure 5.4. Illustration of the extent of haplotype homogeneity in the $-\alpha$ and $\alpha\alpha$ chromosomes observed in the study population.



Genetic distance is indicated at the bottom of the plot in kb. The rows represent each DNA sample tested and the columns represent the SNPs analysed in a 300kb region across the α^+thalassaemia locus represented by the red vertical line. Genes related to the α^+thalassaemia locus are identified by the red box in the lower panel. For each of the markers, the major allele is in black and minor allele in white. The markers are arranged according to the position in chromosome 16 region. It is worth noting that unequal sample sizes of the $-\alpha$ and $\alpha\alpha$ chromosomes were used here. The lower panel of the plot shows genes in the region. Gene names follow the HUGO classification. Gene structure is shown by the blue boxes (exons) and the blue lines (introns). The gene transcription direction is shown by the arrows. All gene positions/locations are referenced to GRCh37.

Figure 5.5. Illustration of the extent of haplotype homogeneity in the $-\alpha$ and $\alpha\alpha$ chromosomes observed in the study population stratified by ethnicity.



(a) Chonyi **(b) Giriama** **(c) Kauma**

Figures (a), (b) and (c) show the haplotype homogeneity in the $-\alpha$ and $\alpha\alpha$ chromosomes observed in the Chonyi, Giriama and Kauma ethnic groups respectively. For further description of the axis and the data, see footnote, Figure 5.4.

5.3.5 Correlation between neighbouring markers with HbS or α^+thalassaemia

A detailed pairwise computation analysis using a correlation coefficient matrix (r) to search for SNPs with close correlation with the HbS or α^+thalassaemia locus was performed. Figure 5.6 shows a correlation matrix plot between HbS (rs334) and the top 10 SNPs located in a 1015.3-kb region on chromosome 11. Pie charts were used to show the correlation for each SNP pair intersection. The colour intensity and size of each circle are proportional to the correlation coefficients, which ranged from -1.0 (a perfect negative correlation) to +1.0 (a perfect positive correlation). A correlation coefficient of zero signifies no association between two SNPs. As proposed by [260], the correlation levels were categorized as follows:

- 0 to 0.2: very weak to negligible correlation
- 0.2 to 0.4: low correlation
- 0.4 to 0.7: moderate correlation
- 0.7 to 0.9: strong correlation
- 0.9 to 1.0: very strong correlation

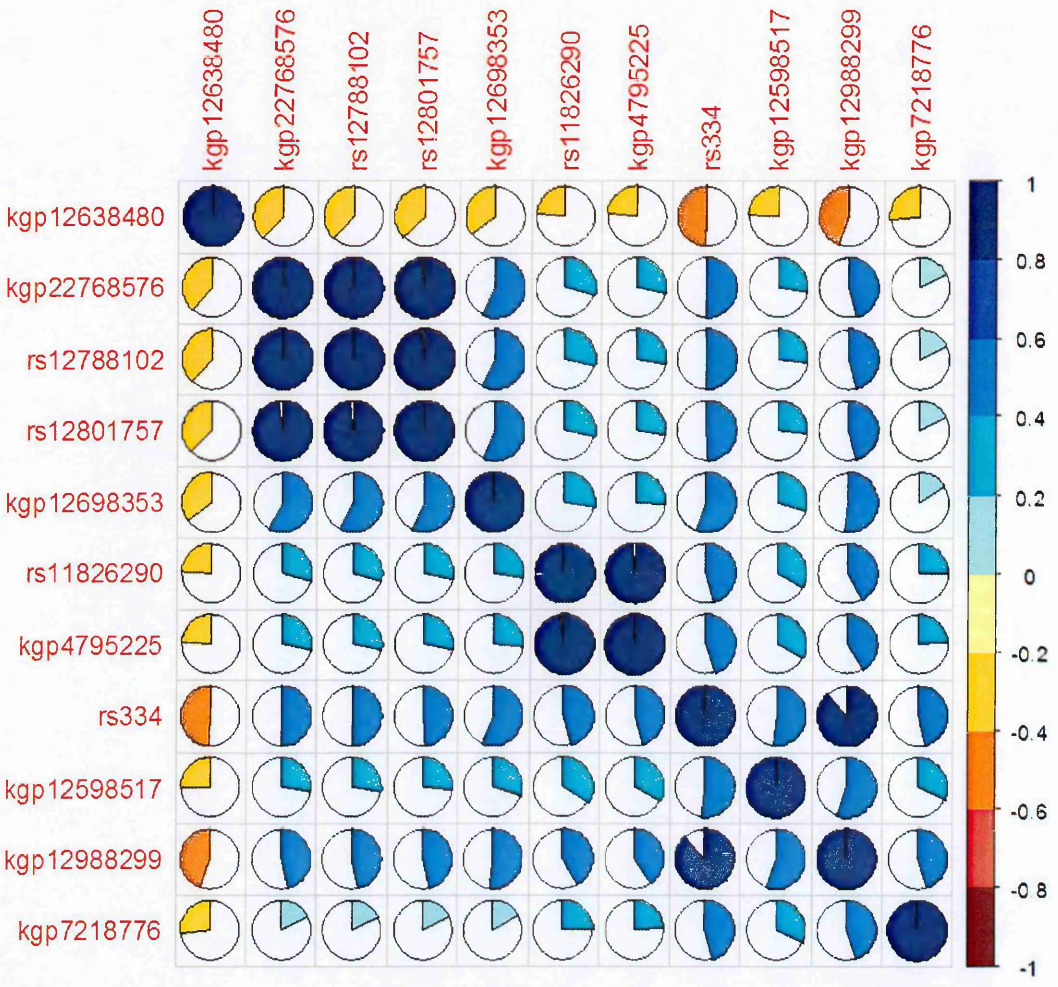
Out of the top ten SNPs, two SNPs: kgp12988299 and kgp12598517 which were 56kb and 15kb upstream and one SNP: kgp12698353 which was 229kb

downstream of HbS were found to be moderately correlated with HbS (Figure 5.6).

Figure 5.7 shows a correlation matrix between α^+thalassaemia locus (thal) and the top 10 SNPs located in a 122.5-kb region on chromosome 16. The interpretation is the same as described above. Out of the top ten SNPs, three SNPs: kgp6975277, rs11248914 and kgp16303049 which were 18kb, 70kb and 87kb upstream of α^+thalassaemia locus were found be low correlated with α^+thalassaemia locus (Figure 5.7).

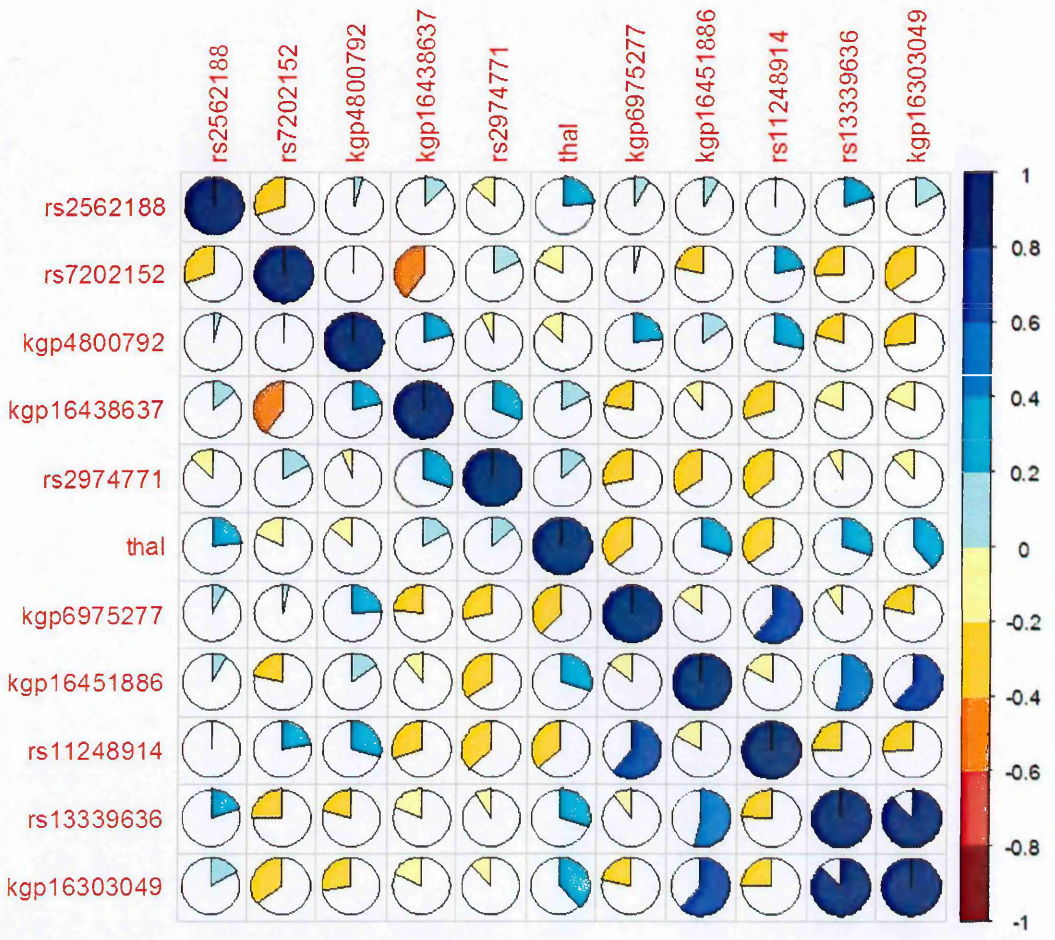
Another useful way for estimating the importance of the r value is to calculate the square of the correlation coefficient (r^2). This squared result gives a rough percentage for the amount of variation between the parameters of interest. On this basis, I also calculated the r^2 values and the results are given in the section that follows.

Figure 5.6. A correlation matrix plot between HbS and top 10 most correlated biallelic markers.



Colour intensity and the size of the circle are proportional to the correlation coefficients. Colour breaks occur at 0.2 intervals. The markers are arranged according to the position in chromosome 11 region. rs334=HbS locus

Figure 5.7. A correlation matrix plot between the α^+thalassaemia locus and top 10 most correlated biallelic markers.

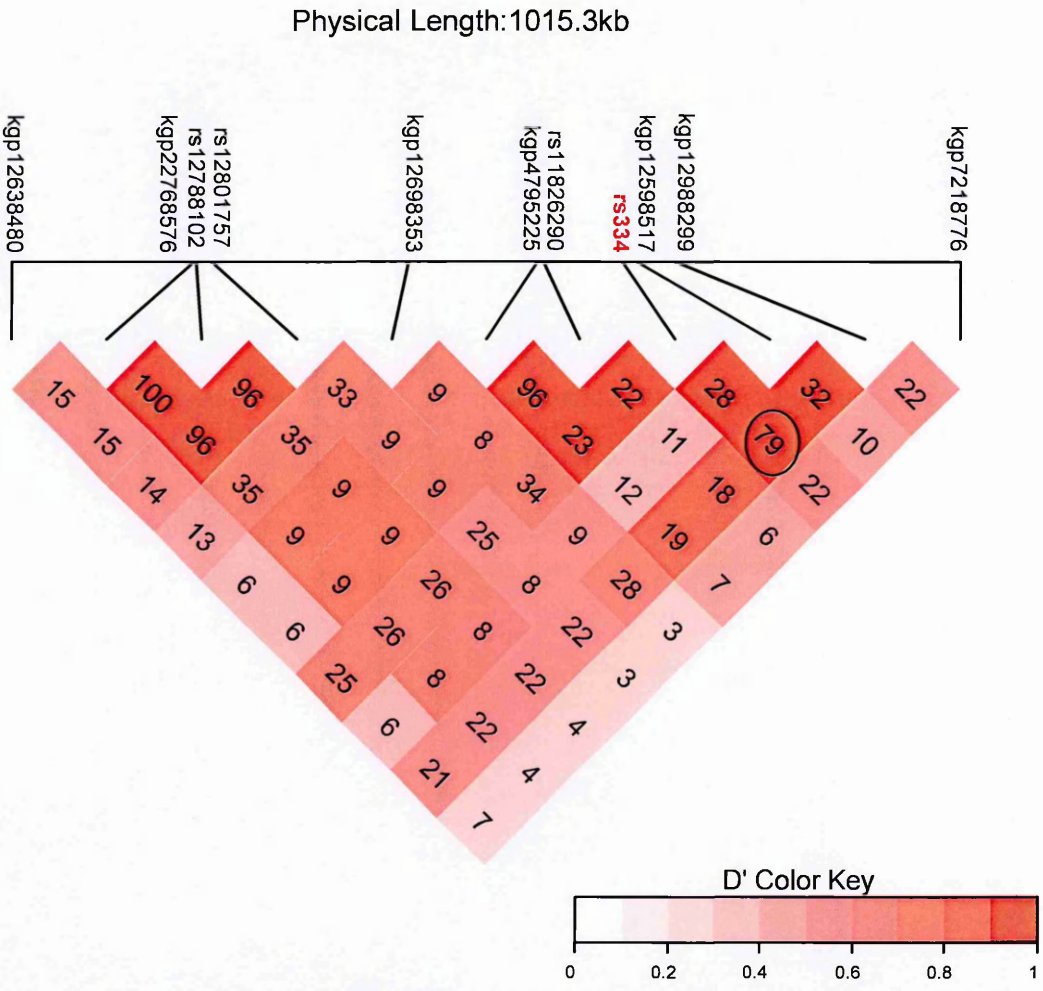


Colour intensity and the size of the circle are proportional to the correlation coefficients. Colour breaks occur at 0.2 intervals. The markers are arranged according to the position in chromosome 16 region. Thal=α^+thalassaemia locus

5.3.6 Linkage disequilibrium patterns

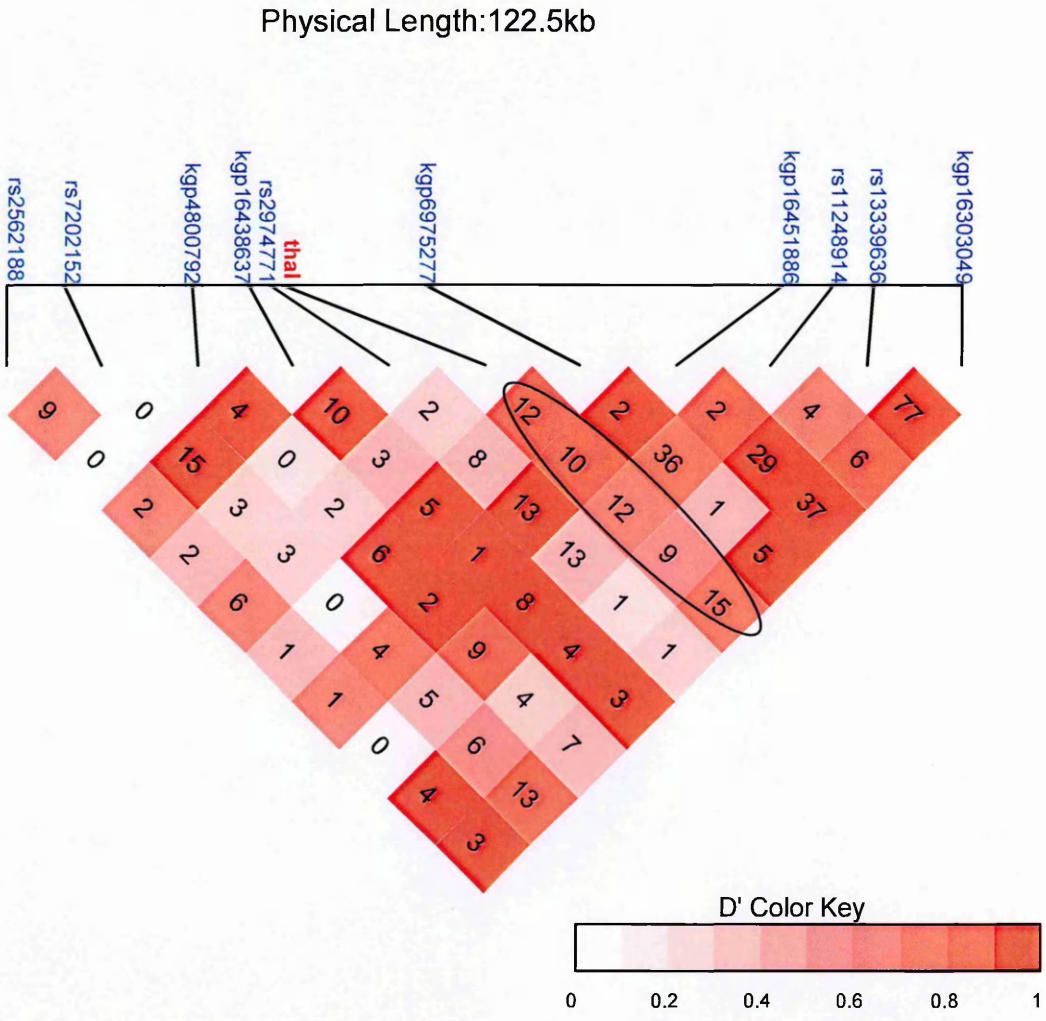
The purpose of a LD map is to tell us whether any two given SNPs are inherited together in an offspring. In other words, we want to know if any given SNP can determine the status of the other SNP. Figure 5.8 provides another view of examining the underlying LD structure among the 10 SNPs surrounding the HbS locus. Here the set of SNPs incorporated into the plot is the same as in Figure 5.6. The markers are arranged according to the position in chromosome 11 region. Red squares represent high pairwise LD, colouring down to white squares of low pairwise LD. D' values from 0.7-1.0 indicate strong LD between pair of SNPs. Whereas D' value <0.7 indicates moderate LD and D' value of <0.2 indicates no LD. D' is 1 when no recombination has occurred between two SNPs. The numbers in the individual squares are the r^2 values multiplied by 100. When the minor alleles at two SNP positions are always present on the same haplotype, r^2=100; when the minor alleles are always on separate haplotypes, r^2=0. A strong LD pattern (r^2=0.79, D'=0.94) was observed between the HbS locus and the kgp12988299 SNP (Figure 5.8 squares marked with a black oval). In contrast, examining the LD patterns between the α^+thalassaemia locus and neighbouring SNPs does not exceed an r^2 of 0.15 (Figure 5.9 squares marked with a black circle), inferring that, no correlation between the α^+thalassaemia locus and neighbouring SNPs

Figure 5.8. The pattern of linkage disequilibrium for ten SNPs surrounding the rs334 locus across a 1015-kb region of chromosome 11.



The markers are arranged according to the position in chromosome 11 region. Red squares represent high pairwise LD, colouring down to white squares of low pairwise LD. D' values from 0.7-1.0 indicate strong LD between pair of SNPs. Whereas D' value <0.7 indicates moderate LD and D' value of <0.2 indicates no LD. The numbers in the individual squares are the r^2 values multiplied by 100.

Figure 5.9. The pattern of linkage disequilibrium for ten SNPs surrounding the α^{+}thalassaemia locus across a 122.5-kb region of chromosome 16.



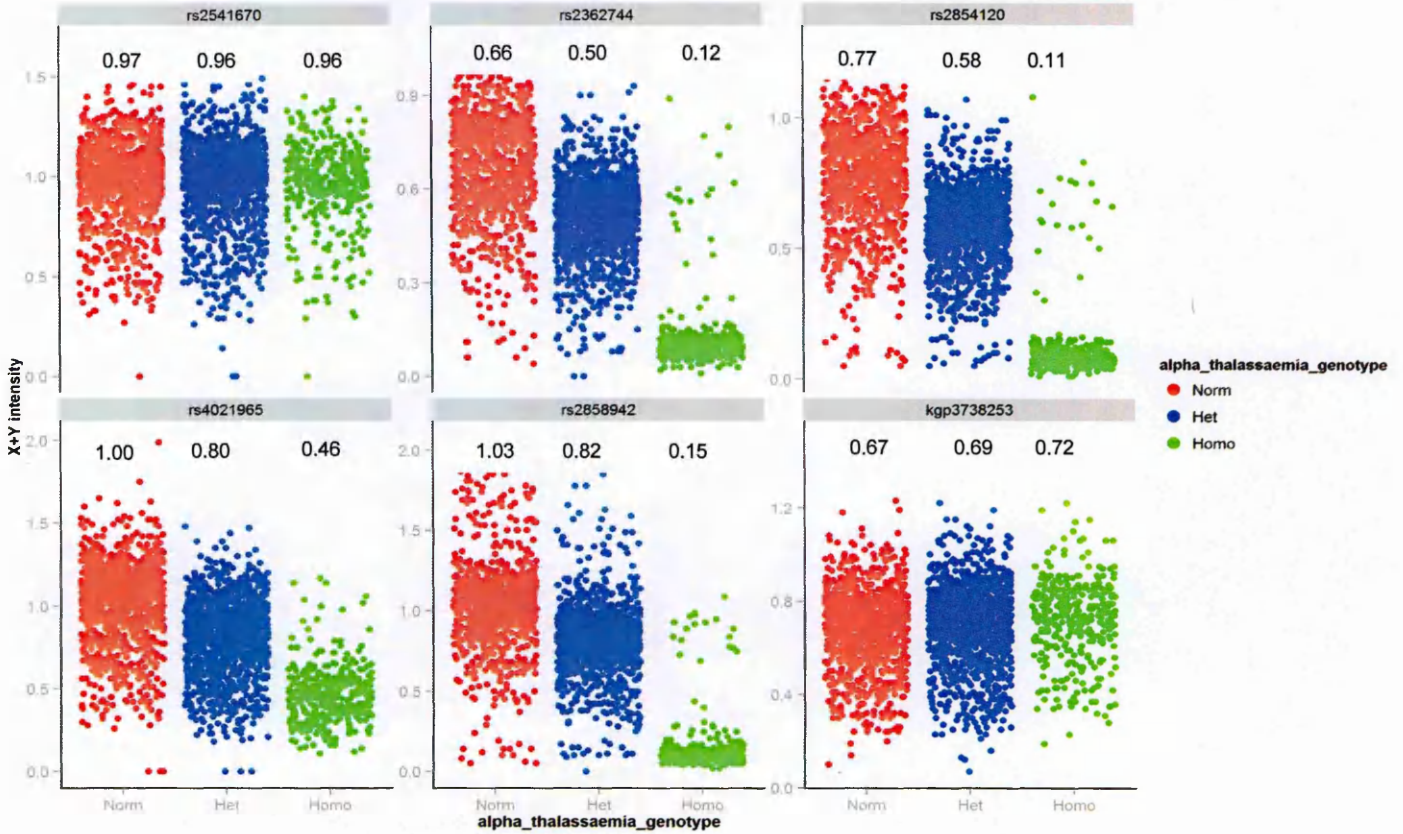
The markers are arranged according to the position in chromosome 16 region. Red squares represent high pairwise LD, colouring down to white squares of low pairwise LD. D' values from 0.7-1.0 indicate strong LD between pair of SNPs. Whereas D' value <0.7 indicates moderate LD and D' value of <0.2 indicates no LD. The numbers in the individual squares are the r^2 values multiplied by 100.

5.3.7 Chip intensity data

From the above analyses it is clear that no apparent haplotype structure was observed in chromosomes carrying the α^+thalassaemia deletion and notwithstanding, no individual marker SNP was found to be strongly correlated or with high LD with the α^+thalassaemia deletion in this population. In the following analysis, I focused primarily on the α^+thalassaemia locus, to examine if using the GWAS Illumina HumanOmni2.5-4 genotyping chip intensity raw data could shed more light on detecting the α^+thalassaemia deletion.

Figure 5.10 is a scatter plot of the normalised intensity of the Illumina HumanOmni2.5-4 genotyping chip by α^+thalassaemia genotype. Each point represents one individual's genotype. The X-axis depicts the α^+thalassaemia genotypes while on the Y-axis are the sums of the SNP X and Y channel intensities. Data for 6 SNPs are shown that span the 3.7kb α^+thalassaemia deletion; rs2541670 and kgp3738253 are outside the deletion and rs2362744, rs2854120, rs4021965 and rs2858942 are internal to the deletion. It can be seen that there is a decrease in overall signal intensities in all α^+thalassaemia homozygotes, although rs4021965 does not show this to the same extent as the other 3. There is also a decrease in intensities in heterozygotes.

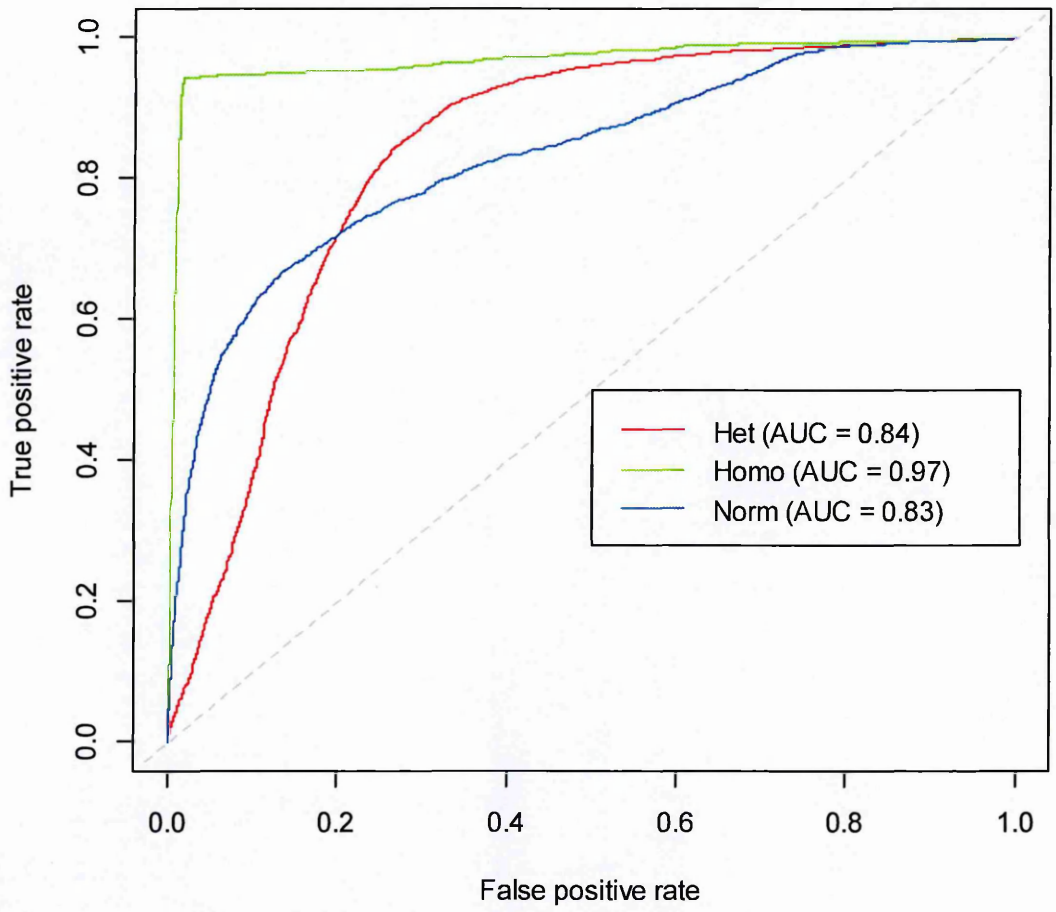
Figure 5.10. A scatter plot of normalised intensity by α^{+}thalassaemia genotypes.



Each point represents one individual's genotype. The X-axis depicts the α^{+}thalassaemia genotypes while on the Y-axis are the combined SNP intensities. The red points denote homozygous wild type, blue for heterozygous and green for homozygous for α^{+}thalassaemia. The four middle SNPs (rs2362744, rs2854120, rs4021965 and rs2858942) are in deletion. The values on top of each genotype are mean intensity values.

In order to test how well the intensity data may be used to determine α^+thalassaemia genotype I have undertaken a combined regression analysis for the 3 SNPs rs2362744, rs2854120, and rs2858942, generated a ROC curve and determined area under the curve (AUC). Figure 5.11 presents a ROC curve showing the combinations of sensitivity (i.e true positive rate) on the vertical axis against and specificity (i.e false positive rate) on the horizontal axis for each possible cut-off value of the continuous chip intensity data that was used to predict the α^+thalassaemia genotype. The results are depicted by bowed curves rising from the 45 degree line to the upper left corner – the sharper the bend and the closer to the upper left corner, the greater the AUC, the greater the accuracy of predictions. From the Figure, we can see that the model was able to identify individuals who were homozygous wild-type (AUC=0.83), heterozygous (AUC=0.84) or homozygous (AUC=0.97) for α^+thalassaemia by using chip intensity data.

Figure 5.11. Receiver operating characteristic curves using chip intensity data with respect to predicting the α^+thalassaemia deletion.



The ROC curve presents the combinations of sensitivity (true positive rate) and specificity (false positive rate) for each possible cut-off value of the continuous chip intensity data that was used to predict the α^+thalassaemia genotypes. Chip intensity data for three SNPs were used: rs2362744, rs2854120 and rs2858942. The dashed reference line represents the ROC curve for a test with no discriminatory ability. The area under the curve (AUC) is displayed in parentheses. The magnitude of the AUC indicates whether the chip intensity data was useful in identifying individuals who were homozygous wild-type (AUC = 0.83), heterozygous (AUC = 0.84) or homozygous (AUC = 0.97) for α^+thalassaemia.

5.4 Discussion

There is growing evidence that haplotype structure may hold the key to understand better human evolutionary history and to identify more efficiently genetic variants underlying complex traits. Linkage disequilibrium has been observed over great physical distances at several genes experiencing recent balancing selection, including common variants implicated in resistance to malaria such as *G6PD* [149, 261-263], *FREM3* [264], *CD40L* [72] and the HbS loci [265]. This has been escalated by the discovery of millions of SNPs. Characterising the haplotype structure and LD patterns of important genomic regions using SNPs is now an important strategy for mapping genes involved in complex human diseases such as malaria. These observations instigated me to investigate how two important variants that confer protection against SM have evolved in the Kilifi population where they co-exist in high frequencies.

The present study investigated the haplotype structure and the LD patterns of two important variants: HbS and α^{+}thalassaemia in the Kilifi population. The results show that the frequencies of these two polymorphisms are indeed different among the three major ethnic groups. For example the frequency of HbSS was higher in Chonyi compared with Giriama and Kauma ethnic groups.

Previous research into the origins of HbS have suggested that this allele may have arisen independently at least five times [266, 267]. The haplotype spans a region of about 60 kb. The distribution of these haplotypes gives rise to their names; Western Africa: Senegal and Benin; Central and Eastern Africa: Cameroon and the Central African Republic (also known as Bantu); and an Arabian-Indian haplotype [256, 268]. I have observed a single HbS haplotype spanning 200kb around the HbS allele which represents the Central African Republic type. However on inspecting the haplotype structure further out to a distance of 1MB on either side of HbS, it can be seen that there are at least 2 major haplotypes and furthermore that these extended haplotypes are differentially represented between the local ethnic groups. Taken together this suggests that HbS arising in Central/Eastern Africa has moved between haplotypes in the local populations but maintained a core region. There are those who have questioned the multiple origins theory of HbS, saying the mutation could have spread to other haplotypes by certain unconfirmed processes, such as recombination and gene conversion [269-271] and therefore propose a uni-centric origin of HbS. My data clearly suggests that HbS can spread to different haplotypes but that it does so with common and large core haplotype in this Kenya group; although there are one or two haplotypes with variation in this core. As further GWAS datasets become publically available

across Africa it will be possible to better explore the multi-centric versus uni-centric origin of HbS.

Conversely, no overall haplotype structure was identified for the α^+thalassaemia locus. However, there did appear to be some structure related to ethnic group. I think that this is due to a combination of several reasons: a very old allele; time for recombination to break all LD down around the locus; time for it to spread onto other haplotypes and then LD to breakdown. Given the homologous nature of sequence covering the HBA genes it is thought that this deletion arises through a problem with homologous recombination. Therefore it is likely that the deletion has arisen several times [272], supporting this hypothesis. If this can happen more than once then it could happen on different haplotype backgrounds with time. Furthermore, the strength of effect on malaria is not large and therefore the length of haplotype is less likely to be long as in HbS.

A strong LD pattern (r^2=0.79, D'=0.94) was observed between the HbS locus and the kgp12988299 SNP. In contrast, no individual SNP was found to be strongly correlated or in high LD with the α^+thalassaemia locus in this population. All this means that imputing α^+thalassaemia locus could be difficult as there are not specific markers tagging the deletion unlike in HbS – although in the latter regional specific tags are important and non-local specific tags will not be good [73].

There are potential limitations in my study. First, the self-reported ethnicity could be creating wrong separation. Nevertheless, this form of classification is the one most probable to be encountered in real-life situations in the clinical practices. In addition, even self-reported ethnicity was able to differentiate clearly groups of individuals with rather different allele (HbSS and homozygous for α^{+}thalassaemia). Second, the study population comprised only Kilifi population and thus the results cannot be freely applied to other populations. Nevertheless, to the best of my knowledge, this is one of the first studies to analyse the haplotype structure of the α^{+}thalassaemia locus including examining the LD patterns and correlation between neighbouring markers. My results should help to refine the genomic structure in both HbS and α^{+}thalassaemia genomic regions and to design case–control association studies involving the Kilifi population. Further studies on various regions across populations will facilitate the proper design of association studies of common complex human diseases. Studying the SNPs in LD is inspired by the observation that some loci are inherited together and may serve similar function.

Chapter 6

A genome-wide scan for loci interacting with known malaria susceptibility genetic variants

Abstract

Genome-wide association studies using the case-control study design have been used widely for finding epistatic interactions in complex diseases; however, to date, there have not been any genome-wide studies searching for epistatic interactions based on an infectious disease such as malaria. Through the MalariaGEN consortium, I have used an imputed GWAS data set that constitutes 9,189,809 SNPs, 1368 SM cases and 1474 controls in Kilifi, to investigate the process of detecting epistasis in SM. I have limited my search space to two well-characterised malaria susceptibility polymorphisms: HbS and α^{+}thalassaemia in a logistic regression framework to search for interacting SNPs/genes. The principles of analysis were demonstrated using one model-pair for each of HbS and α^{+}thalassaemia with the genome yet several interesting findings were made.

One variant, rs4689899 ($P=2.33\times10^{-8}$) belonging to the *STX18* gene on chromosome 4, and 3 variants: rs150797078 ($P=1.75\times10^{-8}$), rs114945726 ($P=2.79\times10^{-8}$) and rs114274411 ($P=2.79\times10^{-8}$) in the vicinity of *MYEOV* gene on chromosome 11 showed compelling evidence for interaction with HbS and α^{+}thalassaemia respectively. This study represents the first GWAS to search for epistatic interactions with respect to SM. Accordingly, my findings here, while requiring further definitive replication, biological and functional validations, highlight some interesting novel interactions.

6.1 Introduction

Genome-wide association studies using SNP markers have identified many common genetic variants associated with common diseases. This has rapidly extended our knowledge of the genetic architecture of these diseases. Nevertheless, compared with the successes of single-locus approaches, the achievements of multi-locus interactions approaches, which seek susceptibilities that derive from gene-gene interactions, have lagged behind [204, 273]. Thus, gene-gene interactions that are largely undetected may explain some of the heritability of common diseases [274]. Most reported interactions related to malaria are currently found through candidate gene approaches, which incorporate prior biological knowledge and no attempt has been made to identify joint genetic effects across the whole genome with malaria as the phenotype of

interest. This is why I have taken advantage of the large amount of genetic data generated through the MalariaGEN consortium to conduct the first genome-wide interaction scan (GWIS) with respect to SM.

Detecting and analysing complex genetic epistatic interactions is not a trivial task, because of the computational difficulty created by the high number of possible interactions for even a relatively small set of candidate polymorphisms. For example, in the candidate gene case-control study in Chapter 4, I was studying ~114 SNPs in 71 malaria candidate genes. These 114 SNPs controlling for confounders yielded more than 50 possible two-pairs interactions. Nowadays a typical GWAS dataset contains many millions of genetic variations, making the enumeration of all pairwise SNP interactions computationally difficult, with the number of tests at hand increasing exponentially. This inevitably becomes a computational bottleneck when it comes to performing a GWIS. Two main strategies to detect epistasis have been proposed in the literature: 1) the conditional search, where at least one locus is fixed, and the other locus is random, and 2) the simultaneous search, which simultaneously scans all possible pairs of interacting loci [275]. Both strategies have been formally examined in linkage [276, 277], and association [236] analysis in humans. Due to the huge search space for complicated multi-locus interactions, I have applied a conditional search methodology.

In Chapter 3, both HbAS and α^+thalassaemia were independently associated with protection against SM as detected through heterozygous and additive models respectively. However, as seen in Chapter 4, instead of resulting in an even greater protective effect, co-inheritance of HbAS and α^+thalassaemia resulted in a reduction in the individual protection afforded by each polymorphism. That said, apart from the interactions seen between HbAS and α^+thalassaemia, are there any other important genetic interactions that might influence associations seen with malaria? For this reason, in this Chapter, I have given priority to these known susceptibility loci because they have been confirmed to be associated with SM risk in my previous analyses, to offer a clear picture on how to detect epistasis in humans using a GWAS study.

6.1.1 Objectives

To date, no full GWAS interaction scans have been undertaken in malaria. As a main goal of this Chapter, I have limited my search space by conditioning on to two well-characterised malaria susceptibility polymorphisms: HbS and α^+thalassaemia and then run these in an interaction scan across the 9,189,809 SNPs genome-wide. This allows the process to be looked into in more detail. To achieve this goal, more specific objectives were set as follow:

1. To perform a QC on the GWIS analysis output through inspecting the distribution of the observed interaction P-values against the expected distribution under the null hypothesis.
2. To analyse any potential signals/hits in more detail including inspecting regions of interest and nature of epistatic effect.

6.2 Methods

6.2.1 Study population

The same subjects and phenotype data from Chapter 5 were analysed in this Chapter. The study samples were collected as part of a genome-wide association study (Kenya GWAS) undertaken by the MalariaGEN consortium (see Chapter 2, section 2.2.3). Some characteristics of the study population are described in Table 6.1.

6.2.2 GWAS genotyping and quality control

Both cases and controls were genotyped for 2,269,360 SNPs at the Wellcome Trust Sanger Institute (more details on genotyping and QC are given in Chapter 2, section 2.2.1.3). SNPs that passed QC were 1,674,680 (74%) SNPs. Imputation using the phase 1 thousand genomes (1000G) reference panel extended the analysis to include 9,189,809 post QC autosomal SNPs. (see Chapter 2 section 2.3.1). All X-linked SNPs were excluded from this analysis.

6.2.3 Statistical analysis

In order to test and demonstrate the process of undertaking a genome-wide interaction scan (GWIS), I ran the analysis with one fixed allele (reference or conditional allele) and used the genome-wide imputation data (>9 million SNPs) as the second allele. Given that for every pair of SNPs there are 25 possible interaction models (5 inheritance models per polymorphism; general, additive, heterozygote, dominant, recessive therefore $5 \times 5 = 25$ possible interaction pairs), I further restricted the analyses to 2 model-pairs. I selected 2 well-characterised protective signals in malaria as the conditional polymorphisms to represent these models (HbS for heterozygous and α^+thalassaemia for additive). The second polymorphism in the interaction scan was used in an additive/general inheritance model. In summary:

- Interaction scan 1:
    - polymorphism A: HbS locus (heterozygous advantage model)
    - polymorphism B: genome-wide SNPs as an additive model
- Interaction scan 2:
    - polymorphism A: α^+thalassaemia locus (additive model)
    - polymorphism B: genome-wide SNPs as an additive model

The test algorithm was SNPepistasis as described in Chapter 2 Section 2.3.2.2. This algorithm was further enhanced by adding a wrapper script by Dr. Gavin

Band (see attributions). I fitted two models; a model that includes an interaction term, and a model that omits an interaction term, and in addition the top five significant principal components (PCAs) to control for population structure were included as covariates in each model.

The GWIS datasets were not submitted in one go, instead it was necessary to split the genome into 'chunks' of approximately 500kb before submitting to a high-performance Linux computing cluster at Wellcome Trust Centre for Human Genetics, Oxford. This was for computational efficiency and also not to over-run the memory capacity of each cluster node. In total 5733 chunks were submitted and one chunk took approximately 1.5 minutes to complete. The data were then subjected to several QC process as detailed below:

1. A Q-Q plot, which shows the distribution of the observed interaction P values of the logistic regression analysis against the expected distribution under the null hypothesis, was generated using the R statistical package (http://www.r-project.org). The genomic inflation factor (denoted λ), was calculated by dividing the median of the observed χ^2 statistics by the median of the theoretical χ^2 distribution [278]. The inflation factor helps in detecting any population structure or genotyping error.
2. A common way of visualising the results of a genome-wide scan is to use a 'Manhattan plot' or a detailed 'regional plot' of a region of interest.

Manhattan plots were generated using the "mhtplot" function of 'GAP', a genetic analysis package for use in R (http://www.inside-r.org/packages/cran/gap/docs/mhtplot) and regional interaction plots were created using LocusZoom ver 1.1 genetic analysis software [279], in which -log10 interaction P-values were plotted against their chromosomal positions.

3. Tables were generated based on the top ten hit regions. Information regarding the SNPs and genes functions included in these tables were sourced from SNP Annotation and Proxy Search (SNAP) [280].
4. Conditional logistic regression was used to assess the independence of interactions, reported for the top most plausible significant hit.
5. The commonly accepted genome wide significant threshold ($P < 5 \times 10^{-8}$) [281], which corresponds to a Bonferroni correction for the estimated one million independent markers was applied to report significant interaction results.
6. To aid interpretation of the results, the nature of the interaction was illustrated by Forrest plots showing genotype combinations. For each combination, an odds ratio from the logistic regression model was calculated by incorporating genotypes and fixed covariates.

6.3 Results

6.3.1 Characteristics of the studied population

The basic demographic characteristics of the study populations are reported in Table 6.1 and are stratified by case–control status. The study included 1368 case patients with severe malaria and 1474 controls. The median age (±standard deviation) of the cases and control children was 27± 23.5 and 7± 2.3 months respectively. Gender was balanced between cases (51% males) and controls (51.3% males). The proportions of SM cases with HbAS, $-\alpha/\alpha\alpha$ thalassaemia, $-\alpha/-\alpha$ thalassaemia were 3%, 47% and 12%, respectively. I performed two conditional GWIS to identify SNPs interacting with two malaria susceptibility polymorphisms: HbS and α^{+}thalassaemia as fixed markers. Each of these scans is discussed in more detail below.

Table 6.1. Demographic characteristics of the study population in the Kenya GWAS stratified by case-control status.

Characteristics		# (%) of cases (N=1368)	# (%) of controls (N=1474)
Age in months, median (SD)†		27± 23.5	7± 2.3
Gender, No. male (%)		702 (51)	757 (51)
HbS genotypes, No. (%)	AA	1307 (96)	1228 (84)
	AS	34 (3)	228 (15)
	SS	17 (1)	11 (1)
α^{+}thalassaemia genotypes, No. (%)	αα/αα	512 (41)	495 (34)
	-α/αα	594 (47)	746 (51)
	-α/-α	154 (12)	233 (15)

†SD, standard deviation.

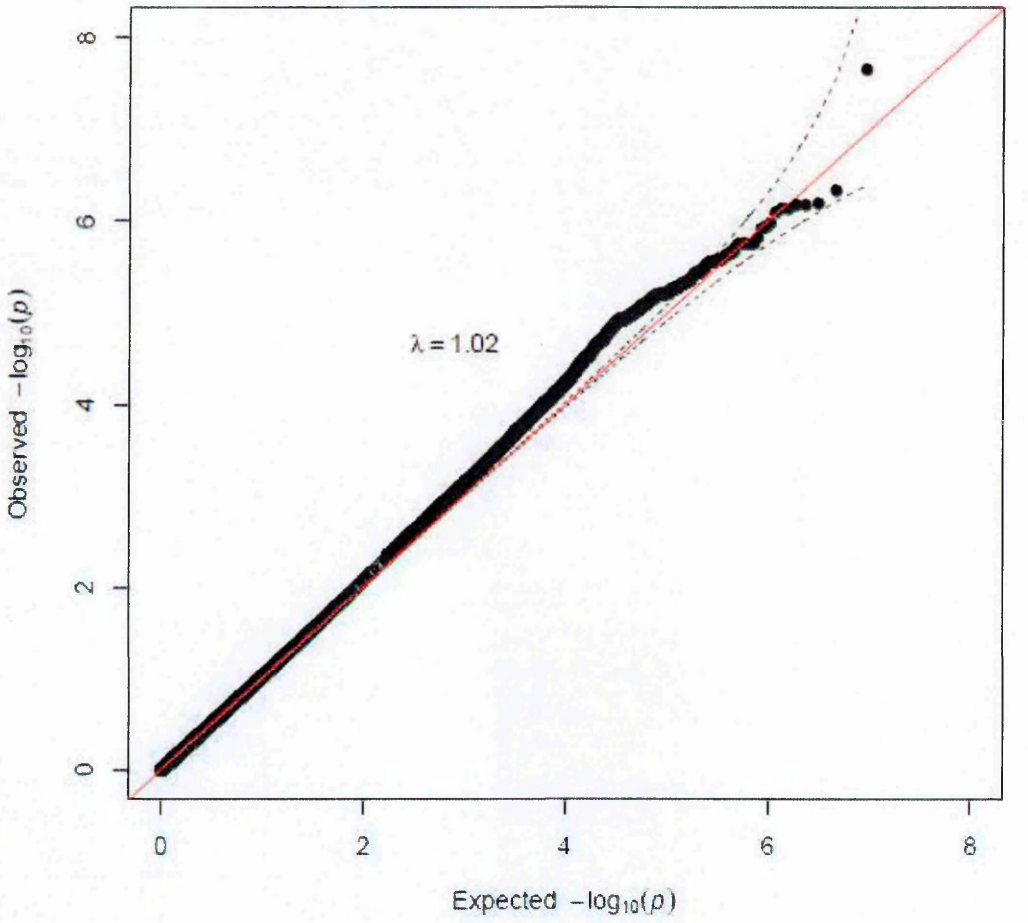
6.3.2 HbS interaction scan across the genome

After exhaustively fitting a conditional two-locus epistatic interaction scan across the genome between HbS as a fixed marker using heterozygous model and all other 9,189,809 SNP markers using additive model, I examined the distribution of the P-values from the interaction term. Figure 6.1 is a Q-Q plot showing that the distribution of interaction term p-values deviates from the expected uniform distribution under the null hypothesis of no epistasis (red diagonal line). These baseline deviations may suggest some interesting results. According to the plot, the genomic inflation factor is 1.02 suggesting that confounding factors such as population structure have been well accounted for in the interaction analysis.

Under those conditions, I visualised the results using a Manhattan plot and plotted the SNPs on the x-axis according to their genomic position on each chromosome against the -log10 P-value of the interaction on the y-axis (Figure 6.2). I observed a strong epistatic effect (interaction P-value=4.42×10^{-8}) between the HbS locus and a marker (rs4689899) on chromosome 4 (the purple diamond) (Figure 6.2). Surprisingly, this was the only SNP to achieve statistical significance at the genome-wide threshold of -log10 P-value of 5×10^{-8} (the solid red line, Figure 6.2). Several other loci (the red diamonds) demonstrated strong but not conclusive interaction evidence with the HbS locus ($P<1\times10^{-6}$) (Figure 6.2). I further investigated the top ten most significant regions showing evidence

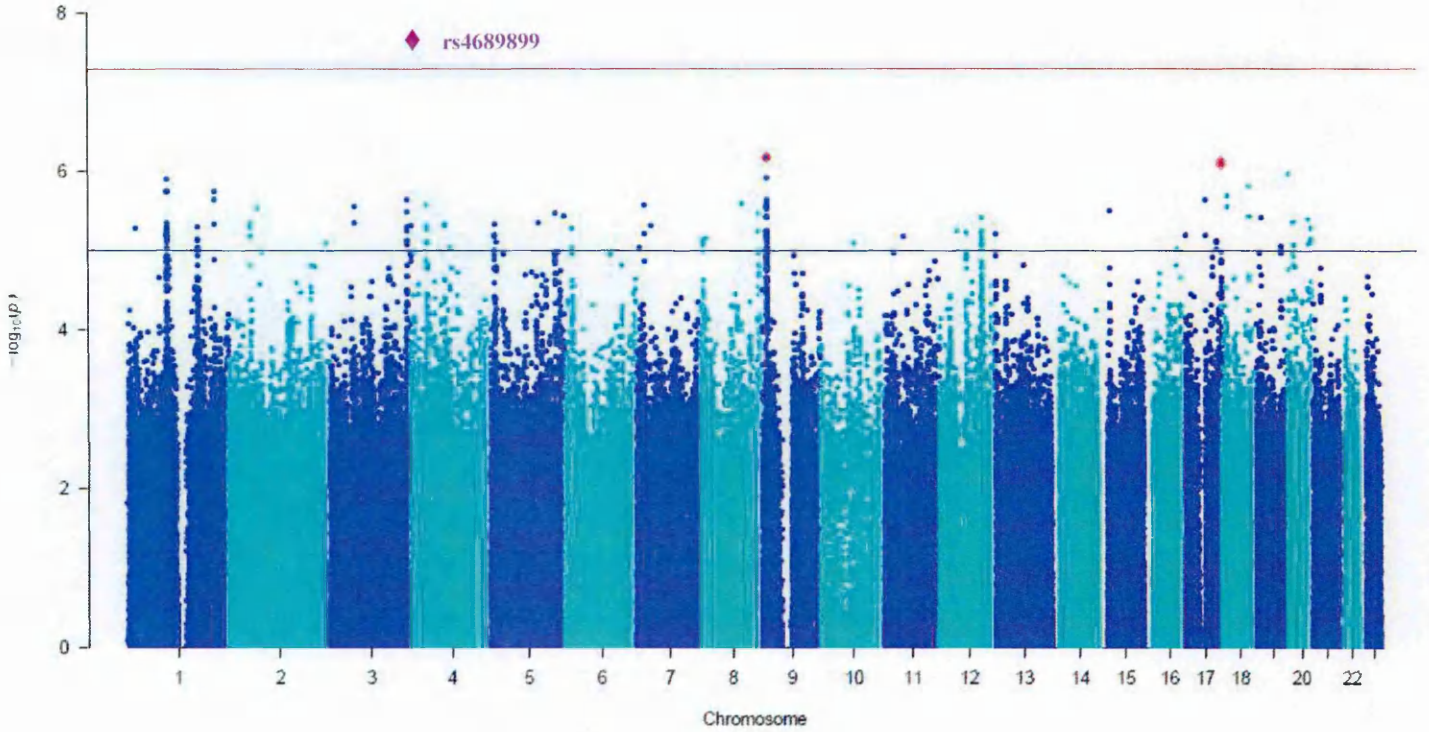
of epistatic effect with the HbS locus (Table 6.2, **Appendix E, and Figure E.1**). When encountering the same gene but with different SNP pair, only the SNP with the smallest interaction P-value was selected. The best interaction signal was the rs4689899 SNP (interaction OR=3.68 (2.32–5.85), interaction P-value=2.23×10^{-8}), located on chromosome 4 within the *STX18* gene (the first region highlighted in blue, Table 6.2). This was followed by rs2586367 SNP (interaction OR=0.24 (0.32–0.85), interaction P-value=6.55×10^{-7}), located on chromosome 9 although not within or near any known genes (the second region highlighted in blue, Table 6.2). The third region, SNP rs57256359 (interaction OR=0.42 (0.22–0.81), interaction P-value=7.26×10^{-7}), located on chromosome 16 within the *JPH3* gene (the third region highlighted in blue, Table 6.2). Moreover, none of these genes have been implicated to have an association with SM. Details on the function and location of the top ten regions showing evidence of interaction with the HbS locus are in Table 6.2 and **Appendix E, Figure E.1**. For purposes of illustration, results presented hereafter are given only for the rs4689899 putative epistatic effect signal.

Figure 6.1. Q-Q plot of the HbS interaction scan across the genome using heterozygous advantage model for HbS and additive model for the interacting SNP.



The plot compares observed −log10 interaction p-values of the tested SNPs on the vertical axis to expected −log10 p-values under the null hypothesis on the horizontal axis. Note the red line was used to indicate a trend that data should follow. Baseline deviations suggested some interesting results. The lambda value indicates the genomic control value for population stratification. The dotted black lines indicate the 95% CI.

Figure 6.2. Genome-wide Manhattan plot showing interaction P-values between the HbS (heterozygous advantage model) locus and interacting SNP (additive model).



P-values are log transformed (-log10) (y-axis) and plotted against chromosomes (x-axis). Chromosomes are coloured alternatively dark and light blue. The red horizontal line indicates the genome-wide significant threshold line at -log10 (5×10^{-8}) while the blue line shows a suggestive threshold at -log10 (1×10^{-5}). The signal in purple diamond above the genome-wide significance is rs4689899. The SNPs coloured in red diamonds are at $P<1\times10^{-6}$.

Table 6.2. Results for the top ten regions showing evidence of interactions with the HbS locus organised by chromosome position.

SNP	Chr	Alleles†	MAF	SNP position	Gene name	Gene description	Gene function	I_{OR} (95% CI)	Int p-value
rs7515285	1	C/T	0.17	99228313	*SNX7*	sorting nexin 7	Involvement in intracellular trafficking.	1.14 (1.18-1.97)	4.42×10^{-6}
rs4689899	4	G/A	0.76	4608405	*STX18*	syntaxin 18	Found to assist with retrograde transport from the Golgi retrograde to the endoplasmic reticulum.	0.68 (1.32-1.85)	2.23×10^{-8}
rs2973697	5	A/C	0.68	177745415	*COL23 A1*	collagen, type XXIII, alpha 1	Still unknown, however, it has been detected at low levels in all connective tissue-producing cells.	2.17 (1.48-3.18)	3.62×10^{-6}
rs2290471	8	A/G	0.46	98943545	*MATN2*	matrilin 2	The specific function of this gene has not yet been determined though GO annotations related to this gene include calcium ion binding.	0.56 (0.37-0.83)	2.52×10^{-6}
rs2586367	9	T/C	0.26	11635003	-		-	0.24 (0.32-0.85)	6.55×10^{-7}
rs57256359	16	A/G	0.16	87699162	*JPH3*	*junctophilin 3*	Provides a structural foundation for functional cross-talk between the cell surface and intracellular calcium release channels.	0.42 (0.22-0.81)	7.26×10^{-7}
rs28409311	16	C/G	0.26	49704798	*ZNF423*	zinc finger protein 423	Acts as a positive regulation of BMP signaling pathway.	1.58 (1.05-2.37)	2.20×10^{-6}
rs7220613	17	C/G	0.59	66319644	*ARSG*	arylsulfatase G	Involved in hormone biosynthesis and modulation of cell signalling.	0.61 (0.41-0.89)	1.52×10^{-6}
rs149664178	19	A/G	0.31	49443450	*DHDH*	dihydrodiol dehydrogenase (dimeric)	Its exact function is not yet known but it is linked to be involved in the oxidation-reduction process.	1.83 (1.26-2.66)	3.97×10^{-6}
rs7247679	19	T/C	0.55	14134172	*RLN3*	relaxin 3	May play a role in hormone activity.	0.85 (0.73-0.96)	4.18×10^{-6}

† reference/derived alleles. For each region the SNP with the strongest signal has been reported. P values were obtained using the LRT. The regions marked in blue were the most significant. Abbreviations: SNP=single nucleotide polymorphisms; Chr=Chromosome; MAF= minor allele frequency; I_{OR}=Interaction Odds ratio; 95%CI= 95% Confidence intervals; Int p-value =interaction p-value.

6.3.2.1 Further exploration of the rs4689899 interaction signal on chromosome 4

According to the HapMap phase 3 recombination maps, the locus of interest , rs4689899 is located in a region of chromosome 4 between two close recombination hotspots, and analysis conducted using the LocusZoom software demonstrated that no long range LD to rs4689899 exist within a 400kb region (Figure 6.3a). Further zooming into a smaller region (60kb) showed that the SNP was actually located inside the *STX18* gene (Figure 6.3b). Among the SNPs in this region, two SNPs showed some evidence of interaction with the HbS, p-value $<10^{-3}$ (**Appendix E, Table E.1**). These two SNPs were rs4689898 that was 32 bases upstream and 4.14kb downstream rs11725796 of the rs4689899 locus. Based on my data the rs11725796 (r^2=0.76) and rs4689898 (r^2=0.49) loci were in moderate LD with rs4689899. To evaluate whether or not the top interaction signal detected in the region was independent from the signal of the above two loci, I performed a conditional analysis. The epistasis interaction between the SNPs and the HbS locus was tested using a logistic regression model with and without conditioning on each one of the SNPs (rs11725796, rs4689898 and rs4689899). Conditioning on either of the SNPs (rs11725796 or rs4689898) evidence for an independent signal was found for rs4689899 locus ($P < 10^{-7}$) (**Appendix E, Table E.1**), suggesting that the strong evidence of interaction observed in rs4689899 locus was independent of rs11725796 and

rs4689898 loci. I also examined the interaction pattern between the HbS and rs4689899 SNP pair using a Forrest plot. Figure 6.4 summarizes the ORs and sample size for each genotype combination relative to the reference genotypes.

In Figure 6.4, we can see that subjects who carry GG genotype for rs4689899 and AS genotype for HbS have an increased disease risk compared to those who carry reference genotypes at both loci (rs4689899:GG/HbAA). However, for subjects who carry GA or AA genotype for rs4689899, carrying AS genotype for HbS significantly decreases the disease risk. For example, if there were no interaction effect, samples that carry GA or AA genotypes for rs4689899 would have an increased risk compared to the reference group. However, they actually have a statistically significantly decreased risk because of the interaction (Figure 6.4).

Figure 6.3. Regional interaction plot for the HbS interaction scan showing the top hit SNP rs4689899 located on Chromosome 4.

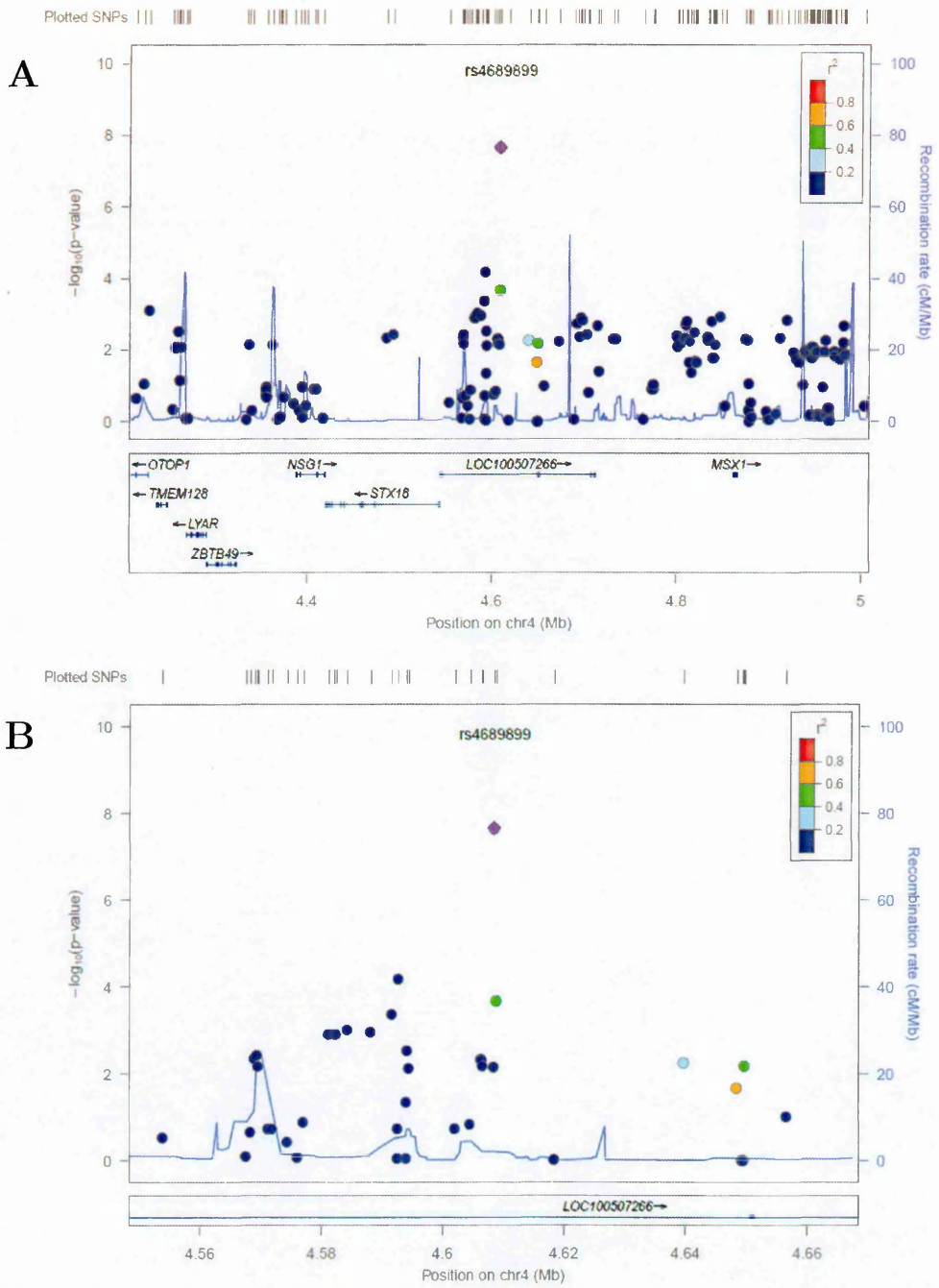
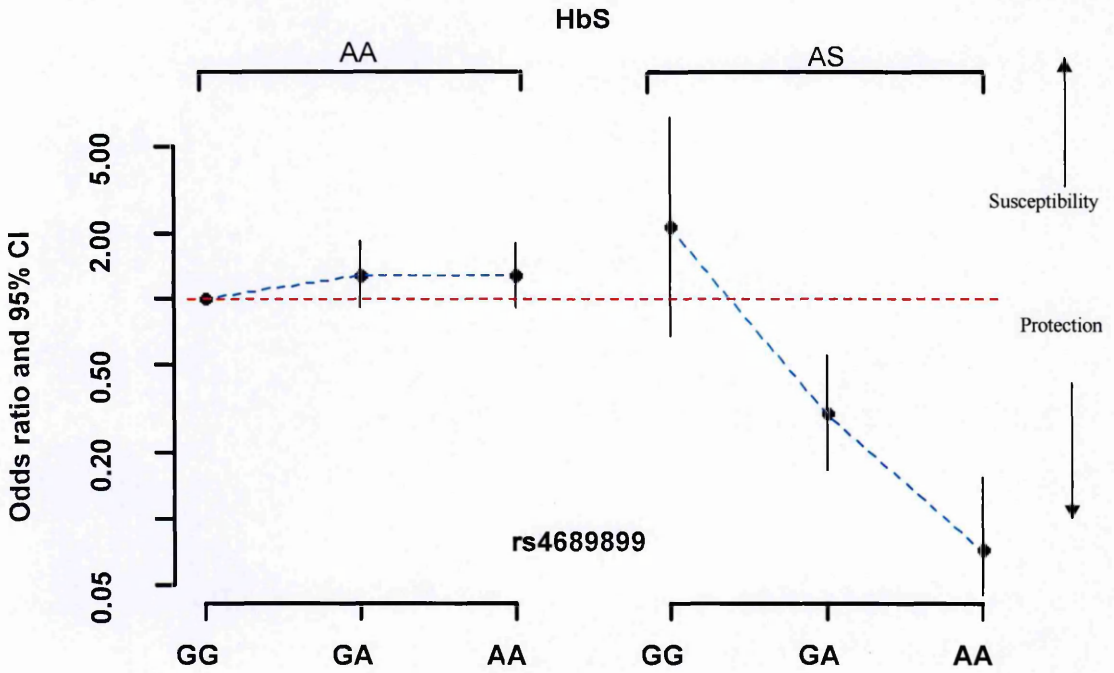


Figure A) shows +/- 400kb around the rs4689899 locus, while Figure B) shows a zoomed in region +/- 60 kb around the rs4689899 locus. The diamond represents the top hit SNP rs4689899. SNPs are coloured based on their linkage disequilibrium (r^2), with the rs4689899 SNP which had the smallest p-value in the region. The –log10 P values for the SNPs are shown in the upper part of each plot. The bottom section of each plot shows the fine scale recombination rates (continuous blue curve along the lower margin of the graph) estimated from individuals in the Hapmap population, and genes are marked by horizontal blue arrows and arrowheads. The plot was produced using locusZoom software.

Figure 6.4. A Forrest plot showing interaction pattern between HbS and rs4689899 SNP-pair.



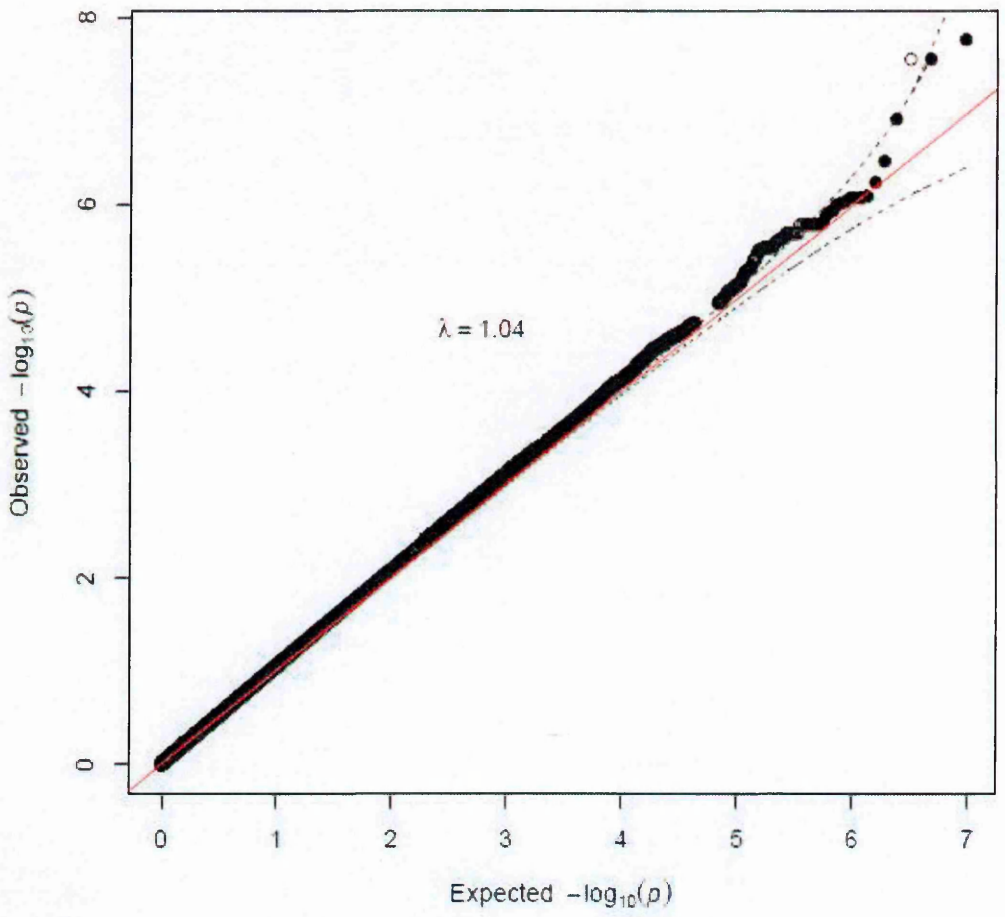
Cases	72	533	836	9	19	9
Controls	82	446	702	5	78	146
P-value	ref	0.19	0.10	0.27	1.64×10^{-5}	7.21×10^{-10}

For each combination of genotypes (x-axis), the odds ratios and95%CIs (y-axis), and p-values relative to the reference group (rs4689899:GG/HbAA) were computed. The red line shows the point of no effect. The odds ratio has been plotted on a log scale so that the distances above and below 1 represent the same size of effect although opposite effect. The sample size for cases and controls are also shown.

6.3.3 Alpha-thalassaemia locus interactions scan across the genome

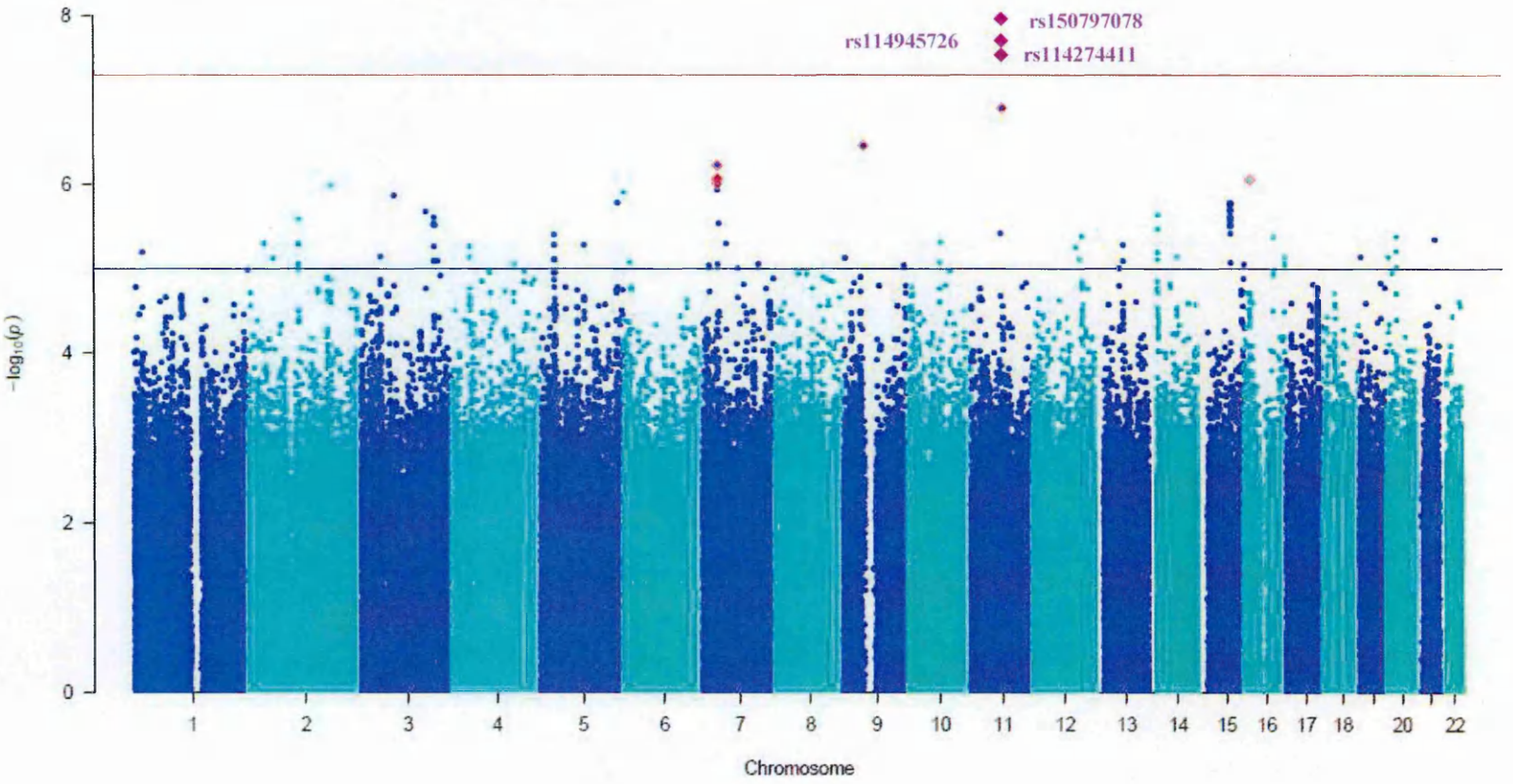
I further performed a GWIS conditional on α^+thalassaemia as fixed effect and the other interacting SNP as random. Figure 6.5 shows the Q-Q plot of observed interaction p-values against those expected under the null hypothesis, indicating no evidence of spurious inflation (λ=1.04). Figure 6.6 shows the Manhattan plot. Examining the epistasis effect between the α^+thalassaemia locus and the rest of the genome, three loci (rs150797078, rs114945726 and rs1142744113), located on chromosome 11 attained genome-wide significance threshold (Figure 6.5). A few other SNPs located on chromosomes 7, 9 and 15 also revealed some evidence of epistasis effect with the α^+thalassaemia locus (points marked in red, Figure 6.5). Results of the top most significant regions showing evidence of interaction with the α^+thalassaemia locus are presented in Table 6.3 and **Appendix E, Figure E.2**, where the SNP id, gene name and its functions, p-value for interaction and as well as local plots of these regions are also listed. The best signals found clustered around *MYEOV* gene with best interaction p-value of 1.75×10^{-8} in rs150797078 (Table 6.3), and 3 more SNPs around the same gene just short of the best interaction p-value. The second region with a peak showing evidence of interaction was found in rs8060012, chromosome 16, base position 12,906,971. This SNP was not within or near any known genes. The third region was SNP rs10225502 located on chromosome 7 within the *AVL9* gene (Table 6.3).

Figure 6.5. Q-Q plot of the α^+thalassaemia interaction scan using an additive model for the α^+thalassaemia locus and the interacting SNP.



The plot compares observed −log10 p-values of the tested SNPs on the vertical axis to expected −log10 interaction p-values under the null hypothesis on the horizontal axis. Note the red line was used to indicate a trend that data should follow. The dotted black lines indicate the 95% CI. The plot shows a slight deviation from the expected distribution (λ=1.04).

Figure 6.6. Manhattan plot summarising the GWAS interaction analysis results for the α^+thalassaemia scan using an additive model for α^+thalassaemia locus and the interacting SNP.



P-values are log transformed (-log10) (y-axis) and plotted against chromosomes (x-axis). Chromosomes are coloured alternatively dark and light blue. The red horizontal line indicates the genome-wide significant threshold line at -log10 (5×10^{-8}), while the blue line shows a suggestive threshold at -log10 (1×10^{-5}). The signals in purple above the genome-wide significance are: rs150797078, rs114945726 and rs114274411. The SNPs coloured in red diamonds at $P<1\times10^{-6}$.

Table 6.3. Results for top ten regions showing evidence of interactions with the α^+thalassaemia locus organised by chromosome position.

SNP	Chr	Alleles †	MAF	SNP position	Gene name	Gene description	Gene function	I_{OR} (95% CI)	Int Pvalue
rs10924388	1	C/A	0.32	246083889	*SMYD3*	SET and MYND domain containing 3	Plays an important role in transcriptional activation as a member of an RNA polymerase complex	1.14 (1.18-1.97)	9.95×10^{-6}
rs7648169	3	A/G	0.48	157864041	*RSRC1*	Arginine/serine-rich coiled-coil 1	Plays a role in pre-mRNA splicing	1.12 (1.03.-1.81)	2.85×10^{-6}
rs4505829	4	G/T	0.65	38418782	*TBC1D1*	TBC1 domain family, member 15	Involved in the trafficking and translocation of GLUT4-containing vesicles and insulin-stimulated glucose uptake into cells	1.36 (1.03.-1.81)	5.41×10^{-6}
rs4479863	5	TG	0.62	167174799	*TENM2*	Teneurin transmembrane protein 2	Involved in neural development, regulating the establishment of proper connectivity within the nervous system	1.70 (1.60.-1.83)	1.62×10^{-6}
rs10225502	7	C/A	0.29	32533950	*AVL9*	-	Involved in exocytic transport from the Golgi	1.40 (1.18.-1.65)	9.60×10^{-7}
rs150797078	11	A/C	0.09	69087089	*MYEOV*	Myeloma overexpressed	Plays a role in gastric cancer cell proliferation and invasion	1.39 (1.04-1.85)	1.75×10^{-8}
rs10861761	12	C/T	0.6	108034827	*BTBD11*	BTB (POZ) domain containing 11	Unknown-integral component of membrane	0.85 (0.72.-0.99)	4.11×10^{-6}
rs8060012	16	G/T	0.37	12906971	-	-	-	0.80 (0.68-0.95)	8.59×10^{-7}
rs3827077	20	C/T	0.33	3721456	*HSPA12B*	Heat shock protein 12B	Protective actions in endothelial cells.	0.70 (0.59.-0.83)	7.18×10^{-6}
rs8131755	21	C/T	0.45	41611966	*DSCAM*	Down syndrome cell adhesion molecule homolog	Negative regulation of cell adhesion	0.79 (0.67.-0.93)	4.47×10^{-6}

† reference/derived alleles. For each region the SNP with the strongest signal has been reported. P values were obtained using the LRT. The regions marked in blue were the most significant. Abbreviations: SNP=single nucleotide polymorphisms; Chr=Chromosome; MAF= minor allele frequency; I_{OR}=Interaction Odds ratio; 95%CI= 95% Confidence intervals; Int p-value =interaction p-value.

6.3.3.1 Further exploration of the rs150797078 interaction signal on chromosome 11

In this section, I will only focus on the rs150797078 locus which presented the best evidence of epistatic interaction with the α^+thalassaemia locus. The rs150797078 is located in a region of chromosome 11 near the highest recombination hotspot in the region, and analysis conducted using the LocusZoom software exhibited some neighbouring SNPs with strong LD (r^2=0.80) with this lead SNP within a 600kb region (Figure 6.3a). Surprisingly, further zooming into a smaller region (60kb) showed that the SNP was actually located outside of the *MYEOV* gene (Figure 6.3b), leading to the question whether the signal was real or false positive. I further examined the interaction pattern for α^+thalassaemia and rs150797078 using a Forrest plot.

Figure 6.8 summarises the interaction ORs and sample size for each genotype combination relative to the reference genotypes. The interaction ORs of rs150797078:AC/α^+thalassaemia:-**α/αα**, rs150797078:CC/α^+thalassaemia:-**α/αα**, rs150797078:AC/ α^+thalassaemia:-**α/-α** and rs150797078:CC/ α^+thalassaemia:-**α/-α** in Figure 6.8 were 1.06, 0.23, 1.56 and 0.47, respectively. This looks like an unusual interaction pattern. However, it is worth noting that the sample size was relatively small when the genotype of rs150797078 was CC and as a result, all ORs estimates for these genotype combinations had large p-values and wide CIs.

Figure 6.7. Regional interaction plot for the α^+thalassaemia interaction scan showing the top hit SNP rs150797078 located on Chromosome 11.

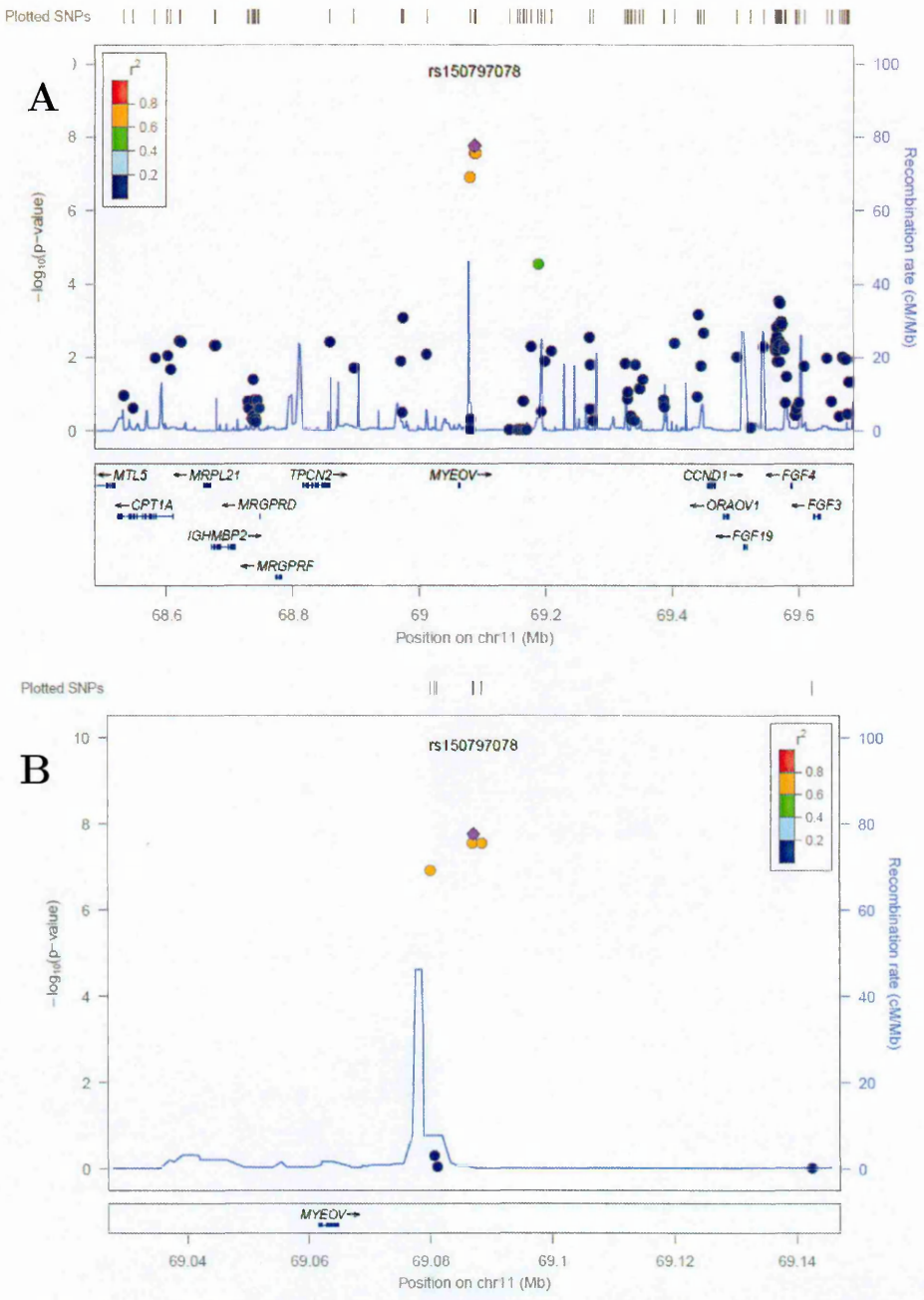
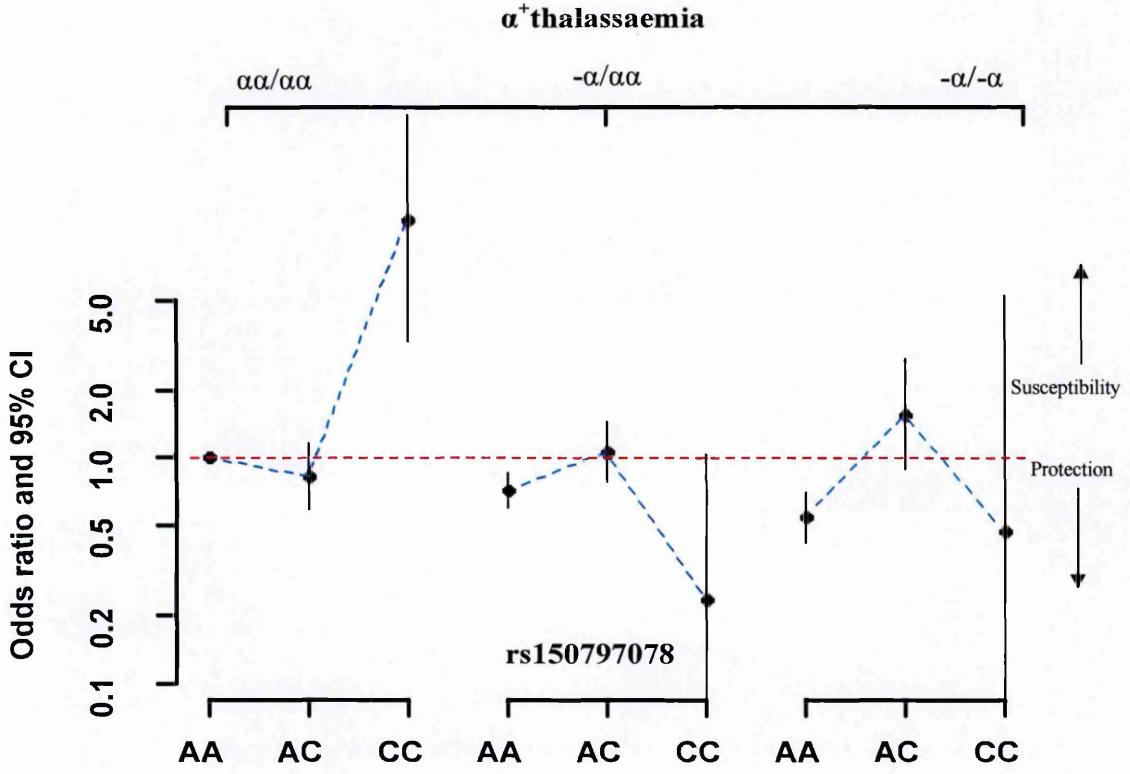


Figure A) shows +/- 600kb around the rs150797078 locus, while Figure B) shows a zoomed in region +/- 60 kb around the rs150797078 locus. The diamond represents the top hit SNP rs4689899. SNPs are coloured based on their linkage disequilibrium (r^2), with the rs150797078 SNP which had the smallest p-value in the region. The –log10 P values for the SNPs are shown in the upper part of each plot. The bottom section of each plot shows the fine scale recombination rates (continuous blue curve along the lower margin of the graph) estimated from individuals in the Hapmap population, and genes are marked by horizontal blue arrows and arrowheads. The plot was produced using locusZoom software.

Figure 6.8. A Forrest plot showing interaction pattern between α^{+}thalassaemia and rs150797078 SNP-pair.



Cases	461	80	10	518	123	2	133	35	2
Controls	409	87	2	635	104	8	212	20	2
P-value	ref	0.26	0.01	2×10^{-4}	0.67	0.05	5×10^{-6}	0.12	0.54

The plot shows the interaction pattern between the two loci for the nine separate genotype combinations. For each combination of genotypes, the odds ratio (95%CI), and p-value relative to the baseline group (α^{+}thalassaemia: αα/αα; rs150797078: AA) were computed the sample size for cases/controls are also shown.

6.4 Discussion

With high throughput genotyping technology, SNP genotyping of a large number of individuals is becoming increasingly practical. Such large-scale SNP genotyping increases the effectiveness of association studies and provides an unprecedented opportunity to study complex genetic effects such as epistasis.

An important role of epistasis in the etiology of complex diseases in humans has been well recognized [247, 282-285]. Although there is appreciation that searching for epistasis in human may be a fruitful endeavour, it still remains a challenging task [274], due to the large number of SNP combinations needed to be tested and the sample sizes of the data sets particularly in the case of GWAS which are enormous. More importantly, the computational difficulty is the main bottleneck; detecting epistasis interaction requires a lot of memory and computational time. The most straightforward approach is to perform a simultaneous search that scans possible pair-wise combinations using some analytic approach. However, in GWAS, this is an intractable computational challenge.

In this large study, I have run a conditional two-locus epistatic interaction scan across the genome each for two malaria susceptibility polymorphisms: HbS and α^{+}thalassaemia as the fixed markers. To check that population structure was

sufficiently accounted for by the inclusion of five PCs, I calculated the genomic inflation factor of the p-values in each scan and found them to all be less than 1.05. The most significant interaction for the HbS scan was observed between the HbS and rs4689899 SNP-pair (interaction $p=2.23\times10^{-8}$). The rs4689899 locus is located on chromosome 4 within the *STX18* gene. This gene encodes an endoplasmic reticulum (ER) protein involved in protein transport between the ER and the Golgi apparatus [286, 287]. *STX18* has also been shown to physically interact with proteins involved in the cell cycle and apoptosis [288]. A GWAS recently conducted by Cordell *et.al* [289], that included 1,995 cases with congenital heart disease and 5,159 controls, identified an association with the STX18 gene and susceptibility to atrial septal defects. The authors further suggested that the possible role of *STX18* involved in regulating cell growth. Based on the available literature, this gene with regard to the risk of SM remains poorly defined.

Among the α^{+}thalassaemia scan, the most promising interaction was between α^{+}thalassaemia and rs150797078 (interaction $p=1.75\times10^{-8}$). However, this SNP occurs at a low-frequency (9%) in my data so these findings should be interpreted with caution. Interestingly, the best 4 SNPs showing evidence of interaction were located on chromosome 11, though nearby the *MYEOV* gene

which plays a role in gastric cancer cell proliferation and invasion [290, 291]. So far, the function of this gene relating to SM has not been clearly defined.

As a starting point to understanding the process of detecting epistasis in humans using a GWAS study, I limited my search strategy to two well-known malaria candidate polymorphisms: HbS and α^+thalassaemia. The reasons of choosing these two polymorphisms were: their prominence in malaria; I was certain about their mode of inheritance for action (i.e. HbS-heterozygous and α^+thalassaemia-additive); existing evidence of their known epistatic interaction that many people believe; and the fact that both polymorphisms co-exist at a high frequency in the Kilifi population. Given the known interaction between HbS and α^+thalassaemia one might expect to see this signal in the two interaction scans that I have performed. However there was no signal in the HBA1/HBA2 region when conditioning on HbS. This is probably due to the fact that, the α^+thalassaemia locus has not been imputed into the dataset or added to the GWAS dataset yet and also, as seen in Chapter 5, no clear markers were in LD with this locus to enable me to detect a signal. On the other hand, it would be more likely to see the interaction for the α^+thalassaemia scan. And indeed I did observe a signal around the HbS locus although the signal was weak ($P<0.08$). This could may have been due to one of the following reasons; first, that there were few individuals in the HbS group leading to a loss of statistical power to

detect the signal; second, that the α+thalassaemia interaction scan was not the correct model (additive × additive) for detecting the HbS signal. The best model would have been additive for the α^{+}thalassaemia locus and heterozygous for the HbS locus. So overall it is not particularly surprising that, I did not observe a stronger interaction association signal between the two loci. .

Although both GWIS pointed to potential epistasis interactions, it is worth pointing out that only two genetic models were applied: for the conditioned loci. For future applications, a model selection technique needs to be developed to determine the most appropriate model with the least loss of power. While not all epistatic effects are likely to follow heterozygous or an additive model, they were chosen since they were the best models associated with protective effect for HbS and α^{+}thalassaemia respectively (see Chapter 3). There is still no agreement in the literature about the most efficient way to perform a GWIS. A number of statistical methods are applicable to the detection of such interactions [204, 292-295] and none of them could be considered the best. Comparing the performances of different methodologies is of great importance but out of the scope of this chapter. In the present study I focused on the application of a standard methodology, the logistic regression model, that has been shown to be a valid methodology for detecting interaction between SNPs [204]. This study had several methodological and technical advantages in addition to the above

interesting findings. (1) To my knowledge this is the first genome-wide study to explore how SNP-SNP interactions influence severe malaria risk. (2) The sample size included in this study is large enough to increase the power of detecting any epistasis interaction if any. (3) My analytical approach embraced inclusion of covariates which provides a more accurate estimate of the interaction effects in SM. (4) These data is part of a GWAS undertaken by the MalariaGEN consortium, indicating that my observed interactions can easily be validated [200]. My study had several limitations. (1) I used the conditional search strategy, where at least one locus was fixed and the other locus was random for the interaction analyses, I could have missed significant interactions between all possible SNPs pairs (i.e. pair-wise comparisons). (2) The small cell size in the interaction analyses might introduce false positives. (3) My approach is data-driven (statistical epistasis), without utilising any existing biological knowledge (e.g., pathways analysis, networks, and other functional annotation data), which may reduce the statistical power and interpretation of my results. That said, the use of this methological approach should be considered as only one step in detecting epistatic interactions in a GWAS. Considerable work still may be required for digesting the test results. Nevertheless, if one can successfully identify a real epistatic interaction; I believe that it will give important clues to the understanding of the underlying biology of human complex traits such as SM.

Chapter 7

Discussion and conclusions

This chapter concludes this thesis, discusses the contributions this thesis makes and considers future directions.

7.1 Statement of contributions

Over the last decade, the study of the genetics of infectious disease susceptibility has undergone substantial revolutionary change [296]. The wealth of information that can be gleaned from the completed map of the human genome coupled with rapid advances in genotyping technologies has presented many opportunities and avenues for scientist to test many long-held assumptions about the pathophysiology of complex human disease such as malaria. This has spawned new areas of specialization in genetics and new methods for the analysis of the genomic data.

In this context, the first contribution of this thesis was in the area of confirming previously proposed associations between malaria candidate genes and SM and its major sub-phenotypes (CM, SMA and RD) by comparing a group of children

with these phenotypes with healthy controls. This line of work was motivated by a large-scale characterised case-control study of severe *P. falciparum* malaria, which contained 121 loci (of these, 65 were previously published associations while 56 were new hits identified in a recent GWAS undertaken by the MalariaGEN consortium) in 71 malaria candidate genes including α^{+}thalassaemia for 2,245 cases and 3,949 healthy controls. This case-control study was carried in Kilifi County, in Kenya, and it was undertaken as part of the MalariaGEN Consortium Project 1 (CP1) [45]. Being part of a consortium that set out standardised phenotypic definitions and high quality genotyping data made the design, assembly and analysis of the study easier.

This is the largest case–control study of malaria that has yet been conducted in Kilifi. This large sample size allowed me to investigate the heterogeneity of the effect of malaria resistance loci in greater detail than has previously been possible. Single SNP association analysis using the 121 polymorphisms including α^{+}thalassaemia revealed twenty-five loci that were strongly associated with SM and its major sub-phenotypes. Out of 71 candidate genes investigated, I observed that polymorphisms affecting various aspects of RBC (including *HBB*, *HBA*, *G6PD*, *FREM3*, *INPP4B*, *ATP2B4* and *ABO*) were among those associated with the strongest signals of differential susceptibility to SM. These genes cover surface proteins that have been shown/suggested to aid parasite

invasion. All these genes, including a new one (ATP2B4) that was recently identified to play a role in calcium ion exchange across the RBC membrane, their mechanisms of malaria protection have not yet been fully elucidated for any; however, with overwhelming evidence that they are important, perhaps this will encourage more RBC related work to understand how humans fight malaria but also lead to new treatments/drugs to fight malaria. This study and a recent report published by the MalariaGEN [45] confirm that most of markers associated with SM are related to RBC and whether there are other markers with the evidence is something worth debating. Some of the RBC markers had strong effects in all phenotypes which is suggestive of a common growth-inhibitory effect. Previous studies have suggested that some of the traditional candidates might be associated with specific effects against particular sub-phenotypes of malaria. In this study however, there was little to suggest that this is true for most of the candidates, including α^{+}thalassaemia which others [16, 41, 62, 66] have previously suggested might be particularly protective against anaemia. The explanation for this discrepancy probably relates to the sample size of these previous studies have been relatively small. In addition, to the traditional candidates, in this study, I have confirmed the associations of *ATP2B4* [230] gene, which is a membrane calcium transport protein, *FREM3* gene, which is an extracellular matrix protein which may play a role in cell

adhesion but little is known on severe malaria and *INPP4B* ,which is a protein binding gene. These are immediate neighbours to the *GYPA*, *GYPB*, *GYPE* region and more likely markers for that region (MalariaGEN personal communication and manuscript under review [213]). These candidates have also been replicated in another MalariaGEN study site in Tanzania [297] and may potentially provide interesting targets for learning about malaria biology and developing new treatments.

Other major polymorphisms of immune mediators such as cytokines and their receptors did not replicate in this study. For example, in my study, no association between SM with *ICAM1* gene was observed. This is in agreement with the results reported by a study in The Gambia, Malawi and Kenya [106]. Such heterogeneity of effect should be kept in mind when judging the lack of replication of many reported SNP associations to disparate malaria clinical manifestations [298, 299].

In a previous study conducted in Kilifi, it has been estimated that inherited factors might account for as much of 35% of the variability in the risk of severe malaria [34]. Typing for a large number of candidate genes gave me the opportunity to re-visit this question using a different approach. I found that the proportion of the total variation in the risk of SM that can be explained by the

polymorphisms that were associated with a significant effect in the current study was only 7.6%, suggesting that additional genes are yet to be discovered. This was derived from an ascertained set of SNPs and with the current GWAS's under analysis this can be estimated across the whole genome across several MalariaGEN sites.

Part of the motivation for the recent enthusiasm for looking for genetic interactions underlying the human disease is the sense that previous failures to identify and especially to replicate, significant individual genetic associations might be driven by underlying complexity generated by epistasis [283]. Given this rapid increase in the size and precision of human association studies, we are now entering an era in which we can rigorously address the hypothesis of whether failure to replicate these associations was due to genetic complexity. The second contribution of this thesis was therefore in the area of investigating whether the malaria candidate genes discussed above were acting independently or whether some of them interact to bring about protection. Most of the reported epistasis between malaria candidate genes has been investigated in small studies using data on a limited number of genes. To my knowledge, this is the first large-scale case-control study of SM to investigate epistasis between multiple candidate genes. In the present study, I investigated epistasis by applying three computational algorithms: PLINK, AntEpiSeeker and

SNPepistasis. I compared and contrasted the performance of these three approaches. All the three methods selected the same topmost hit SNP pair, and additional significant epistatic interactions were observed. While the three methods confirmed the previously reported epistatic interaction between HbAS and α^{+}thalassaemia the strength of association was rather weak, perhaps for reasons of study design. The combination of the strong protective effect of HbAS and the relative rarity of compound heterozygous individuals meant that even a study of this size included few individuals in the combined groups, resulting in limited power.

Searching and reporting epistasis results involve two major challenges. First, since an enormous number of possible combinations are tested, a large proportion of significant associations might be false positives. In this case, I validated my results using STRING, a database of known and predicted protein-protein interactions [245]. By using this established biological database, I might miss novel epistasis interactions between genes; nevertheless the interaction models with detectable statistical epistasis, having enough biological evidence has a high likelihood of being interpretable. The second major challenge involves interpretation of the results and developing them into biologically meaningful hypotheses. For this reason, I focussed on HbAS and α^{+}thalassaemia epistasis interaction that I believe to be true because it has

been repeated in multiple populations from Kenya, Ghana and Mali [57, 62, 103]. In the light of this observation, as a third contribution of this thesis I investigated the dynamics of evolutionary selection of HbS and α^+thalassaemia genomic region. It is possible that an apparent association with malaria can arise from LD between the typed SNP and a primarily associated polymorphism. Constructing a detailed haplotype structure map of polymorphism around these candidate genes and describing a profile of LD among them will be helpful to identify the causal variants(s). Also since neither HbS nor α^+thalassaemia are typed on the current GWAS chips having a useful marker on the chips in LD could help determine the HbS and α^+thalassaemia genotypes savings us time, effort and cost on direct genotyping especially if the assay is non-trivial. To display the haplotype structure, I decided only to show homozygotes as the data has been phased and was possible that the heterozygous individuals for these loci have not been assembled correctly. I compared patterns of haplotype homogeneity in the chromosomes that were homozygous for HbS (HbSS) and homozygous for α^+thalassaemia ($-\alpha/-\alpha$) with those of homozygous wild type for HbS (HbAA) and homozygous wild type for α^+thalassaemia ($\alpha\alpha/\alpha\alpha$) respectively.

The results of the study indicate that in Kilifi, the HbS allele shows clear evidence of recent positive selection. It lies on haplotypes that are significantly

less diverse than those for the wild type HbA allele. The signature of recent positive selection observed in the HbS haplotype raises critical questions on how they can be explored in doing the predictions, and the effect of other mutations in the HbS haplotype on its anti-malarial properties and disease intensity in different climatic conditions in Africa. The differences in the HbS haplotypic backgrounds observed here is also a reminder of the unresolved debate on the origins of HbS mutation. Conversely, no apparent structure was seen in either the $\alpha\alpha$ or $-\alpha$ chromosomes across the 300-kb region. Nevertheless, some differences were observed in the $-\alpha$ chromosomes by ethnicity. These findings suggested that the $-\alpha$ allele might be an ancient allele that has undergone a breakdown in haplotype structure by recombination or that the mechanism of deletion due to the homology in the sequences across the HBA region has occurred many times. Further, exploring the LD patterns showed a moderate level of LD with the HbS locus, but no individual SNP was found to be correlated or in high LD with the α^+thalassaemia locus in this population. Based on this observation, I extended my analysis to look at the raw chip intensity data to try and predict the α^+thalassaemia genotypes. Given that α^+thalassaemia is an important associated gene in malaria then any GWAS should find a way to type it to allow a full catalogue of genes to be included. Also it will allow further work including epistasis.

Finally, since the central theme of this thesis has been epistasis between malaria candidate genes, I further explored and illustrated how to search for and evaluate epistatic interactions in a GWAS study. Although there is appreciation that searching for epistasis in humans may be a fruitful endeavour, it still remains a challenging task [274], due to the large number of SNP combinations needed to be tested and the sample sizes of the datasets particularly in the context of GWAS studies. To deal with these challenges, as a starting point, I limited my search space to the HbS and α^{+}thalassaemia polymorphisms in a logistic regression framework using one genotypic model for each pair. The study involved 9,189,809 SNPs, 1368 SM cases and 1474 healthy controls from Kilifi population. The most significant interaction for the HbS scan was found between HbS and the rs4689899 locus within the *STX18* gene, an endoplasmic reticulum (ER) protein involved in membrane trafficking between the ER and Golgi, on chromosome 4. Conversely, the most promising interaction in the α^{+}thalassaemia scan was with the rs150797078 locus located on chromosome 11, close to the *MYEOV* gene that plays a role in gastric cancer cell proliferation and invasion. So far, the role of these two genes (*STX18* and *MYEOV*), if any, with regard to the risk of SM remain poorly defined. Furthermore, I have only explored a small snapshot of the epistasis space to understand the process of detecting epistatic interactions using data from a large scale genome-wide study. Apart from

validating and replicating the results, there is more work to be done including pair-wise interactions of potential signals and also performing a full genome wide interaction scan taking into account all the models of inheritance.

This study had some limitations. First, the sampling was not representative across all ages. The majority of the data were from young children under the age of 5 years; reflective of the general burden of malaria disease among young children in the study area. Second, while this study identified some interesting interactions, it must be emphasized that they have not yet been validated and this could be a major limitation to drawing strong inferences about the epistatic interactions observed and whether they are real.

7.2 Future directions

Although the work presented here is a detailed interrogation of malaria candidate genes, there are still avenues that have been left unexplored. The single-SNP association analysis depicted some intriguing findings. The study has power to look at the effects on mortality and parasite densities on the protection afforded by the malaria candidate genes included in the study. Another interesting observation was the proportion of variance explained by these candidate genes additively on malaria risk. It will be interesting to see the

proportion explained by including their interactive effects and extend it to GWAS.

Considering further work relating to epistasis detection, there are multiple folds: First, epistasis models used in this study were limited to heterozygous and additive effect for the genome scan. The full epistasis space still needs to be explored for the other genetic models of inheritance. Second, the obvious next step for these analyses is to link the network structure revealed by epistasis analysis to information obtained from other methods, such as biological network and build a comprehensive map of the full interactions. Given the computational challenge, rather than conducting a full genome wide interaction scan, concentrating on a set of genes known to be functional in protection from malaria, offers a promising way forward. For example, Collins *et. al* [300] used all pair-wise interactions of 743 genes known to influence chromosomal processes such as DNA repair, transcription regulation, and chromatid segregation in yeast. This enabled them to place their epistatic interaction results within the context of already well-explored systems. In this respect, their epistatic interactions corresponded well to known physical interactions among proteins, nonetheless allowed novel interactions to be discovered above what would otherwise be a chaotic set of more than half a million potential epistatic interactions. In my future analysis, I will concentrate on RBC polymorphisms

and examine their joint effects on SM and other major sub-phenotypes. Third, in order to validate some of the observed epistasis, it is my long-term goal to collaborate with the MalariaGEN consortium for further experimental verification and further biological interpretations.

Finally, since this is the first study to characterise the haplotype structure and LD patterns of the HbS and α^{+}thalassaemia locus in Kenya it would be more fascinating to investigate the age of these two alleles, purposely to throw more light on how they have evolved in the Kilifi population. Studying haplotype structures and the chip intensity data are also important as they may identify markers or ways to identify some of the more complex markers that are implicated in malaria. These are not necessarily represented on GWAS chips; deletions such as α^{+}thalassaemia, *GYPC*, *SLC4A1* (Band3), all of these are important to validate and should be annotated within GWAS studies to be able to fully investigate the genetic contribution to malaria resistance.

Appendices

Appendix A

Supplementary Table for Chapter 1

Table A.1. Presents advantages and disadvantages of the epistasis methods available in the literature. The algorithms in blue are discussed in detail in Chapter 2.

Algorithm	Advantages	Disadvantages
FastANOVA, COE	• Offers Several Tests Statistical convex. • Readily available for use.	• Only considers homozygous genotypes; • Requires a small sample size. • Lacks validation, which could result in biased results.
TEAM	• Offers Several Tests Statistical convex. • Readily available for use.	• Lacks validation, which could result in biased results.
MDR	• Good power to detect association and assumes a genetic model priori • Well evaluated and understood by several researchers. • Readily available for use.	• Algorithm for 134,135 samples of genomic scale. • Has problems to detect associations in presence of locus with heterogeneity.
FastEpistasis	• Provides parallel processing module epistasis of Plink; • Scale linearly with the number of processors considered.	• Error in calculating the Variance. • Does not perform estimation of genotypes due to incomplete data.
PLINK	• Good power to detect association assuming a certain type of genetic model on the data. • Widely distributed and available for use.	• The comprehensive test is computationally expensive.
AntEpiSeeker	• Treatable for data of scale genomics; • Method simple and easy to implement • Readily available for use.	• Detection of association in the absence of data with marginal effect weak is lost due to an incomplete search of the space of possible associations.
SNPRuler	• The algorithm is based on learning of rules • Provides easy interpretation. • Do not assume a priori distribution on the data. • Provides a List of Interactions sorted by significance.	• Cannot detect interactions containing combined rules • Does consider models of genetic heterogeneity • Performs No validation to avoid spurious results to reduce false positives.

SNPHarvester	• Complexity of linear search. • Provides the possibility of removing SNPs with significant marginal effect for the correct detection of interactions.	• The removal of SNPs with marginal effect significant limits the possibility of identifying all results of interactions • The random selection of the initial set of SNPs may limit the detection of significant associations.
BEAM	• Allows incorporating expert knowledge using a priori distribution on the data. • Good power to detect association in interaction models with MAF low • Readily available for use.	• Problems to detect associations on data of genotypes without main effect • Execution time is slow in comparison with other methods.
Epiforest	• Random Forest are Quick to build • Good power to detect interactions with pure epsitasis effect. • Supports multiple file formats.	• Problems to detect interactions with little or no marginal effect. • Uses a consensus vote which limits the list of susceptibility loci with the phenotype under study.

To determine the state of art and put my own work in context, I compared twelve common approaches that have made the search for epistasis a computational reality and enumerated their advantages and disadvantages.

Appendix A

Appendix B

SNPepistasis -Pseudocode for Chapter 2

B.1 SNPepistasis – pseudocode.

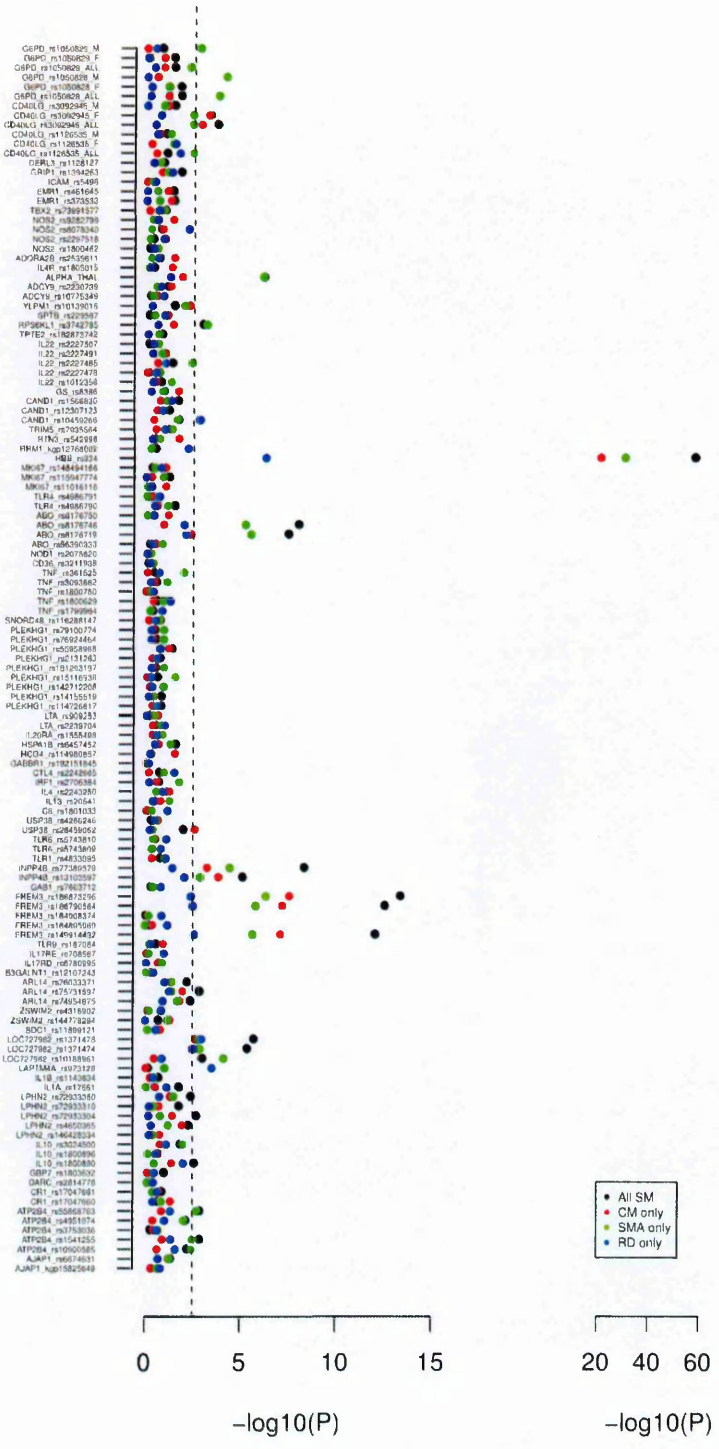
```
#############################################################
####################FUNCTIONS################################
#############################################################

options(help_type="text")
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}
fNumeric=function(temp) as.numeric(as.vector(temp))

fSNPdataModel=function(snp,model="general"){
 if(model=="general"){
  snp=as.factor(snp)
 }
 if(model=="additive"){
  snp=as.integer(snp)
 }
 if(model=="het"){
    snp[snp!=1]=0
 }
 if(model=="dominant"){
  snp[snp!=0]=1
 }
 if(model=="recessive"){
  snp[snp<2]=0
  snp[snp==2]=1
 }
 snp
}
## set working directory ###########################################
Home="/data/bayes/users/cndila/2015/Method2/"
setwd(home)
## Read data with all the 16 models ################################
modelSpec=read.table("Models.txt", header=T)
modelSpec[,2]=as.character(modelSpec[,2])
modelSpec[,3]=as.character(modelSpec[,3])
no.modelSpec=dim(modelSpec)[1]

## Read data the genotype data and fit the
regression#########################
db=read.csv("Data_2014.csv", header=T)
output.file="output.txt"
outputMinimum.file="outputMin.txt"
c.range=dim(db)[2]
Case=db$CASE
D=db[,-c.range]; #D=D[,1:5]
snps.names = names(D)
n.snps = length(snps.names)
snps.names.l1 = snps.names[-n.snps]
iCount=0
result = data.frame();result.min = data.frame()
for(snp1.name in snps.names.l1) {
     iCount=iCount+1
```

```
snp1=D[,snp1.name]
cat('Running interaction scan of ', snp1.name, ' (i.e. ', iCount, ' of ', n.snps, ') with other
snps...\n',sep="")
j = which(snp1.name==snps.names)+1
for(snp2.name in snps.names[j:n.snps]) {
snp2 = D[,snp2.name]
modelSNPs=paste(snp1.name,'*',snp2.name,sep="")
this.result.min = cbind(ModelSNPs=modelSNPs,
          count_00=length(which(snp1==0 & snp2==0)),
          count_01=length(which(snp1==0 & snp2==1)),
          count_02=length(which(snp1==0 & snp2==2)),
          count_10=length(which(snp1==1 & snp2==0)),
          count_11=length(which(snp1==1 & snp2==1)),
          count_12=length(which(snp1==1 & snp2==2)),
          count_20=length(which(snp1==2 & snp2==0)),
          count_21=length(which(snp1==2 & snp2==1)),
          count_22=length(which(snp1==2 & snp2==2)),
          interaction.P.min=NA
                    ) # end cbind
                    print(snp2.name)
        this.result=NULL
        for (modelNo in 1:no.modelSpec){  #for modelNo
          modelSpec1=modelSpec[modelNo,2]
          modelSpec2=modelSpec[modelNo,3]
          snp1M=fSNPdataModel(snp1,model=modelSpec1)
          snp2M=fSNPdataModel(snp2,model=modelSpec2)
          modelSpecs=paste(modelSpec1,'_',modelSpec2,sep="")
                    tryCatch({
           g=glm(Case~snp1M*snp2M,family='binomial')
                     d=drop1(g,scope=g$formula,test='LRT')
           this.result=rbind(this.result, cbind(ModelSNPs=modelSNPs,ModelSpecs=modelSpecs,
          mPval1= d['snp1M', 'Pr(>Chi)'],mPval2= d['snp2M', 'Pr(>Chi)'],
          iPval= d['snp1M:snp2M','Pr(>Chi)']))
          },error=function(e){print(e);}) # end tryCatch
        } # end for modelNo
        this.result=data.frame(this.result[,1:2],apply(this.result[,3:5],2,fNumeric))
        index.min=which(this.result$iPval==min(this.result$iPval))
                    result.min = rbind(result.min,this.result[index.min,])
                    result = rbind(result,this.result)
        } # end of snp2.name for
} # end of snp1.name for
## Saving minimum pvalues of the models ##################################
cat( 'Saving results to \"', outputMinimum.file, '\"...\n', sep = "" )
write.csv(result.min, file = output_minimum.file, row.names = F, quote = F )
## Saving the all the results ############################################
cat( 'Saving results to \"', output.file, '\"...\n', sep = "" )
write.csv(result, file = SNPepistasis_output.file, row.names = F, quote = F )
## Viewing the top 100 most significant pvalues ##########################
cat( 'Complete.  Top 100 most significant interaction P-values:\n', sep="")
options(width=300)
result.min= result.min[order(as.numeric.factor(result.min['iPval']), decreasing = F ), ]
print( result.min[1:100,] )
cat( "bye!\n" )
```

Appendix C

Supplementary Figures for Chapter 3

Figure C.1. Shows the distribution of minimum p-values from the genotypic tests of inheritance.

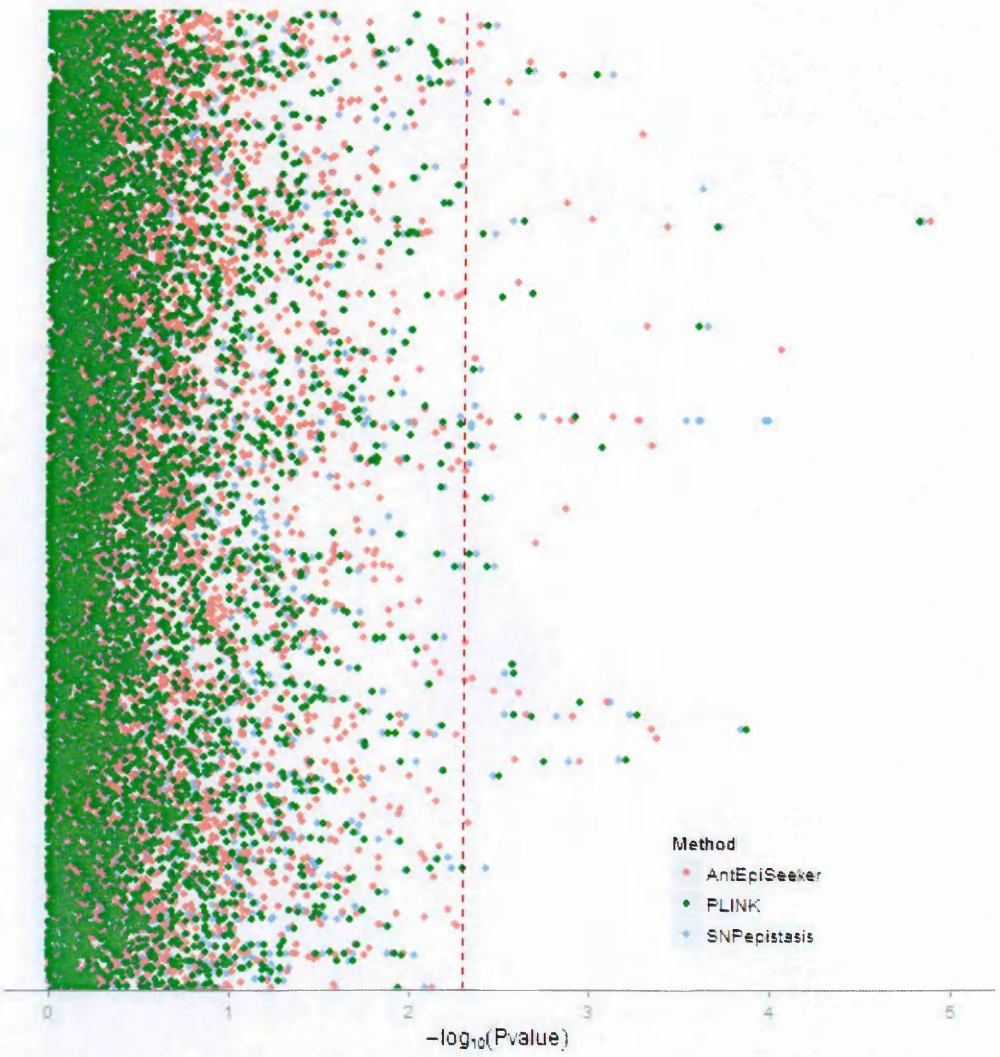


Minimum P-values derived from additive, dominant, recessive and heterozygous advantage models, adjusted for gender, HbS and ethnicity for All SM and sub-phenotypes; CM only, SMA only and RD only applied to the 122 SNPs including α^{+}thalassaemia. The dashed line represents a p-value threshold of <0.005.

Appendix D

Supplementary Figure for Chapter 4

Figure D.1. A Manhattan plot for all pairs epistatic interactions detection using the 3 methods.



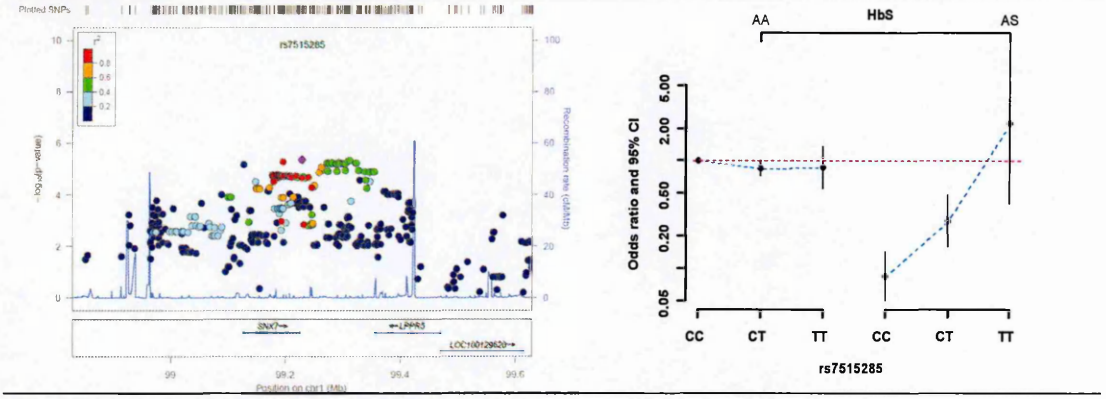
The dashed line represents a p-value threshold of <0.003. The X-axis depicts –log p-value and the Y-axis are the SNPs pairs.

Appendix E

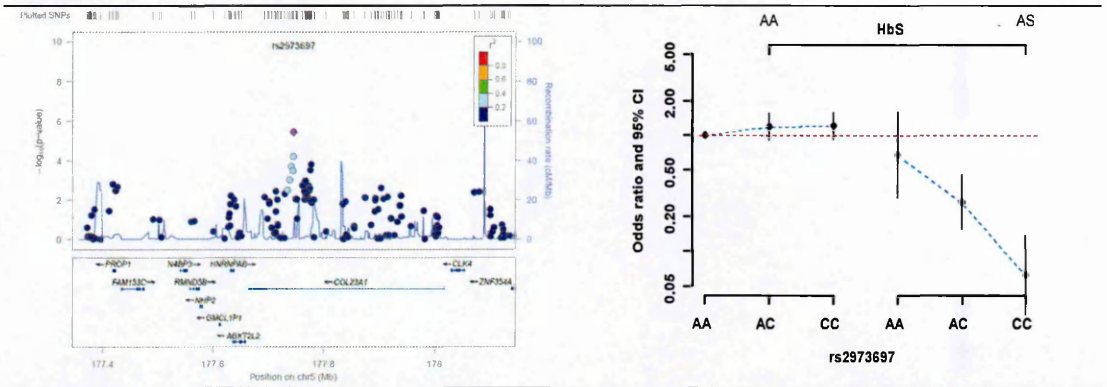
Supplementary Figures and Tables for Chapter 6

Figure E.1 The top ten significant regions of the HbS interaction scan across the genome.

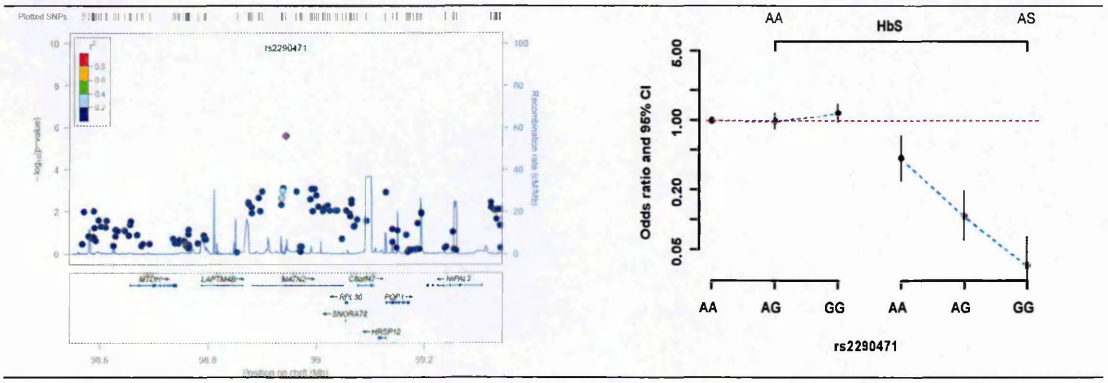
Region 1. A regional plot and a Forrest plot of interaction between HbS and rs7515285 on chromosome 1.



Region 2. A regional plot and a Forrest plot of interaction between HbS and rs2973697 on chromosome 5.

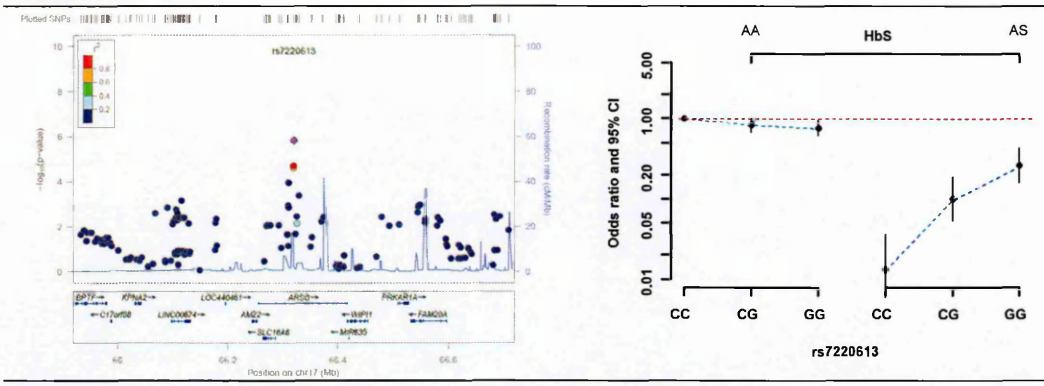


Region 3. A regional plot and a Forrest plot of interaction between HbS and rs2290471 on chromosome 8.

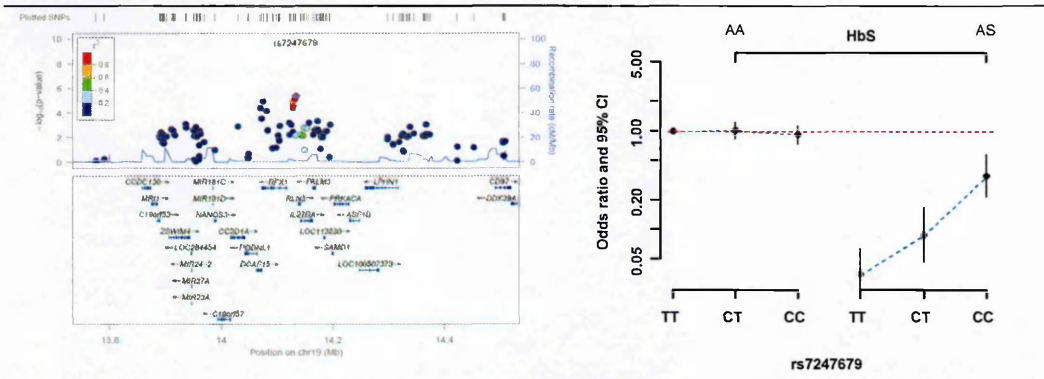


The left panel shows +/- 400kb around the top hit SNP, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio +95%CI (y-axis). The red line shows the point of no effect.

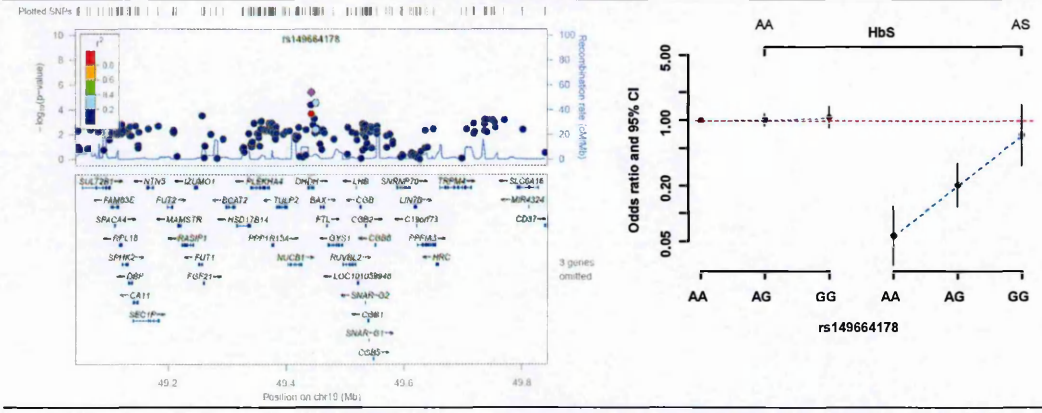
Region 4. A regional plot and a Forrest plot of interaction between HbS and rs7220613 on chromosome 17.



Region 5. A regional plot and a Forrest plot of interaction between HbS and rs7247679 on chromosome 19.

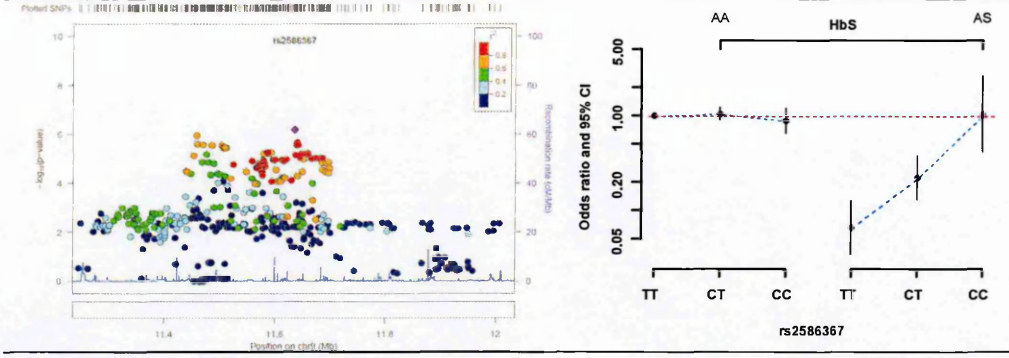


Region 6. A regional plot and a Forrest plot of interaction between HbS and rs149664178 on chromosome 19.

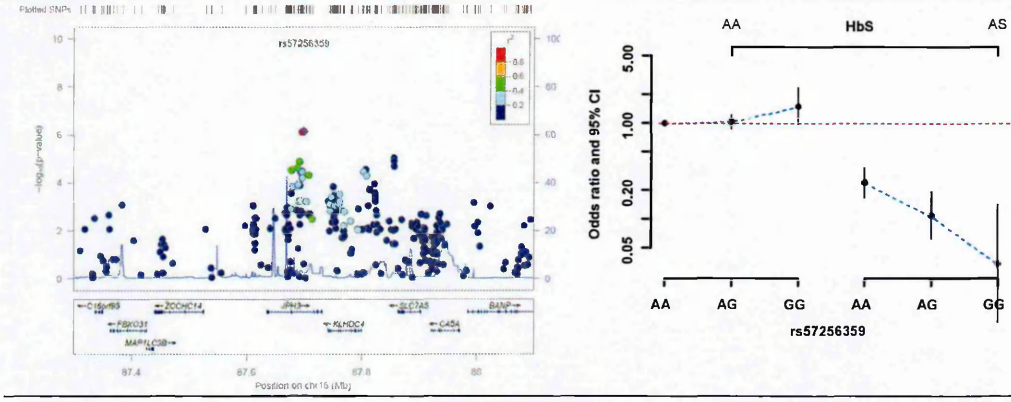


The left panel shows +/- 400kb around the top hit SNP, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio ±95%CI (y-axis). The red line shows the point of no effect.

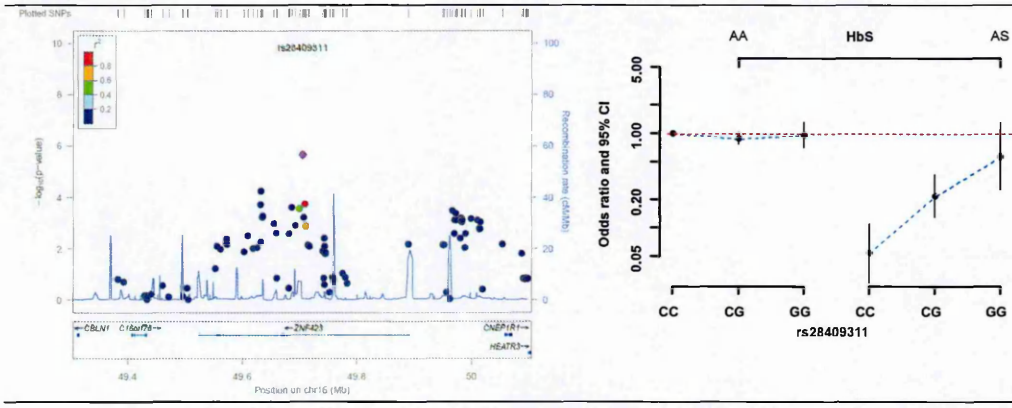
Region 7. A regional plot and a Forrest plot of interaction between HbS and rs2586367 on chromosome 9.



Region 8. A regional plot and a Forrest plot of interaction between HbS and rs57256359 on chromosome 16.



Region 9. A regional plot and a Forrest plot of interaction between HbS and rs28409311 on chromosome 16.



The left panel shows +/- 400kb around the top hit SNP, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio ±95%CI (y-axis). The red line shows the point of no effect.

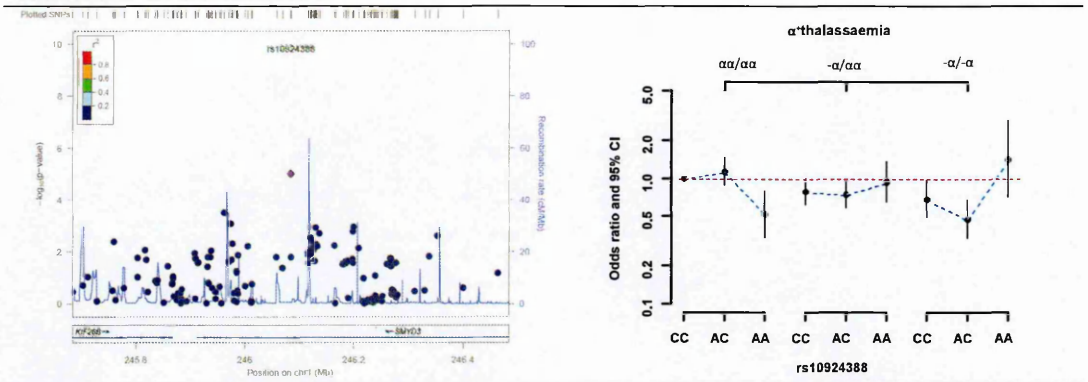
Table E.1. Conditional logistic regression was used to assess the independence of interactions, reported for the top most plausible significant hit located on *STX18* gene region with the HbS locus.

CHR	SNP A	SNP B	r^2	SNP A Interaction P	SNP B Interaction P	SNP A given SNP B Interaction P	SNP B given SNP A Interaction P
4	rs11725796	rs4689899	0.76	2.07×10^{-4}	2.23×10^{-8}	7.00×10^{-4}	2.24×10^{-8}
4	rs4689898	rs4689899	0.49	6.99×10^{-3}	2.23×10^{-8}	1.00×10^{-3}	2.28×10^{-7}

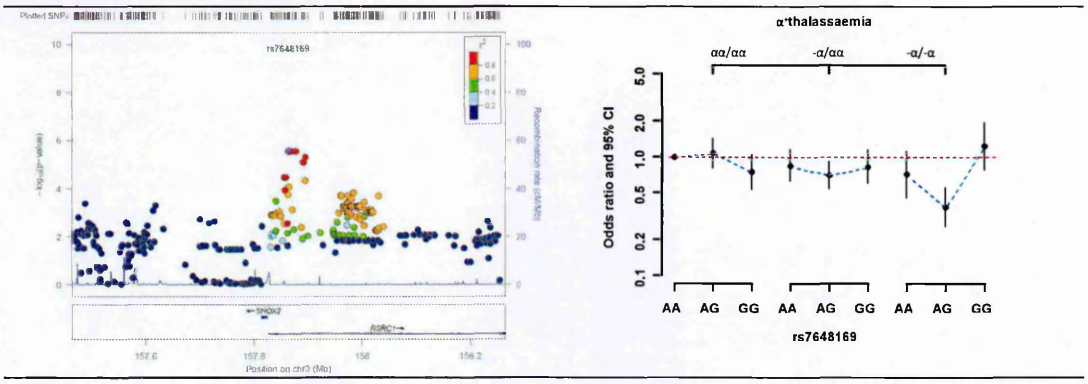
P-values were calculated using LRT and adjusting for the first five principal components as covariates. Abbreviations: Chr=Chromosome; SNP A =the neighbouring SNP near the top most hit; SNP B is the top most significant hit in the HbS GWIS; r^2= the LD between SNP A and SNP B; Interaction P values are results for epistatic interaction between SNP A or SNP B with the HbS.

Figure E.2. The top ten significant regions of α^+thalassaemia interaction scan.

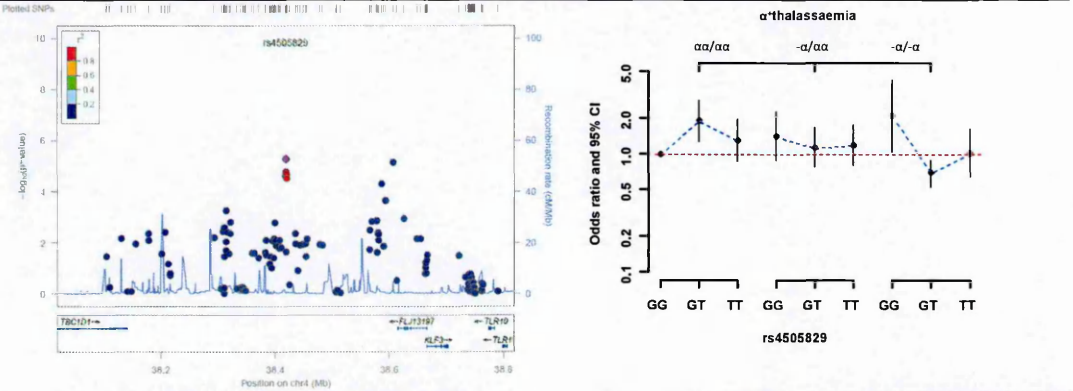
Region 1. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs10924388 on chromosome 1.



Region 2. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs7648169 on chromosome 3.

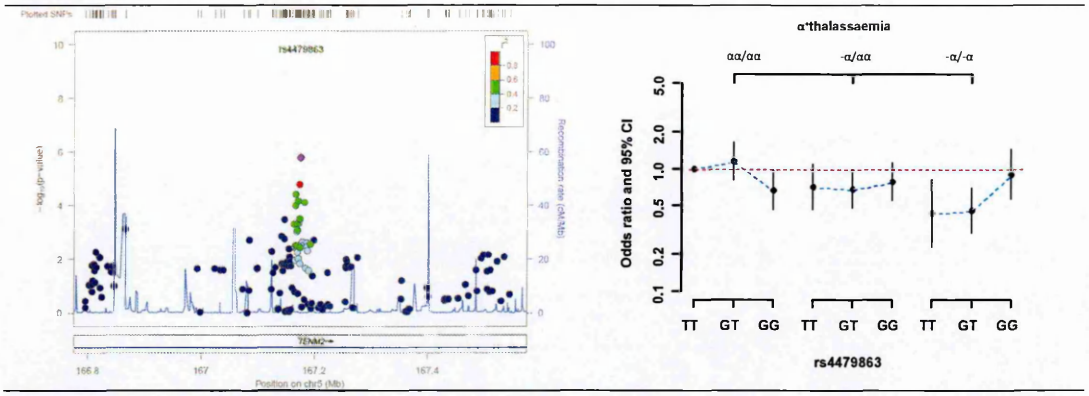


Region 3. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs4505829 on chromosome 4.

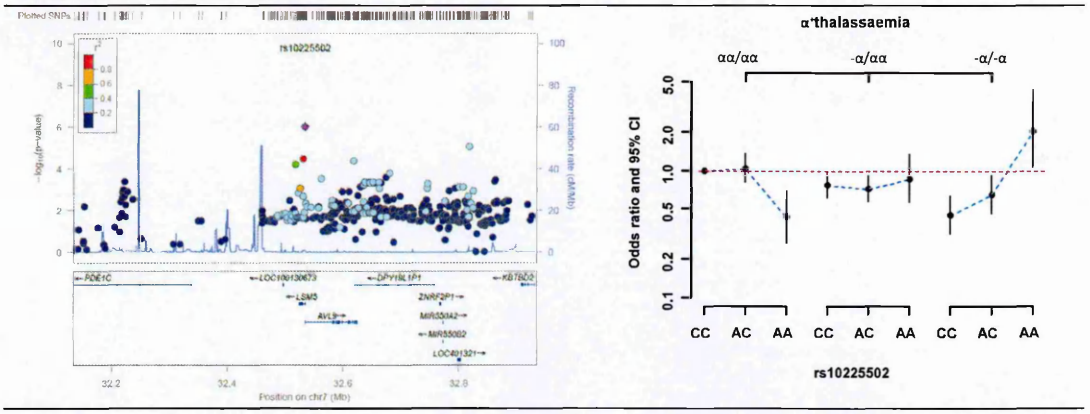


The left panel shows +/- 400kb around the top hit SNP for the α^+thalassaemia GWIS, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio ±95%CI (y-axis). The red line shows the point of no effect.

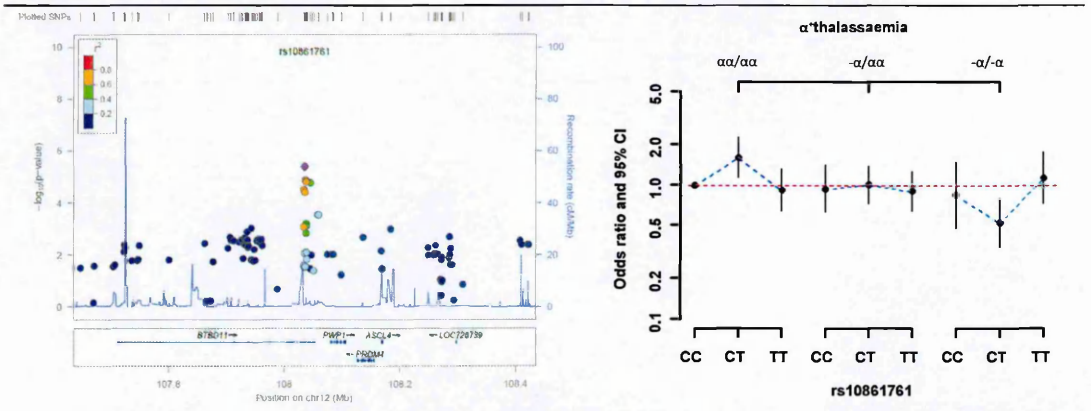
Region 4: A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs4479863 on chromosome 5.



Region 5. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs10225502 on chromosome 7.

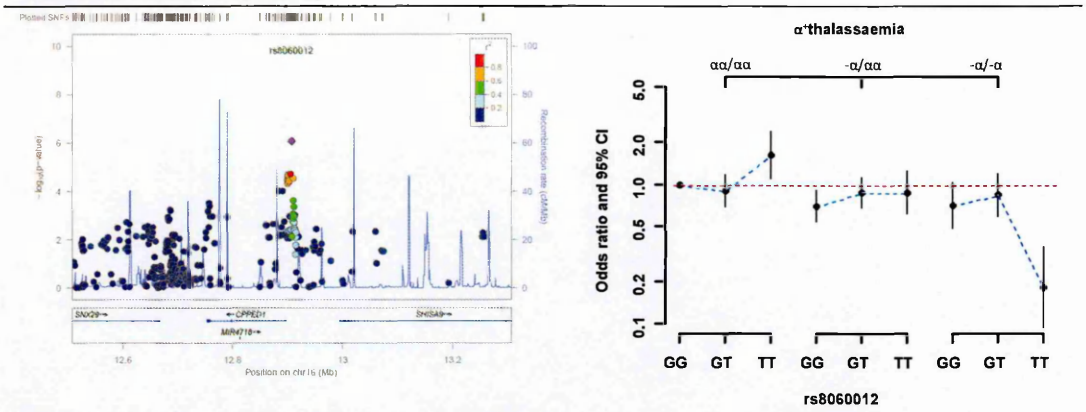


Region 6. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs10861761 on chromosome 12.

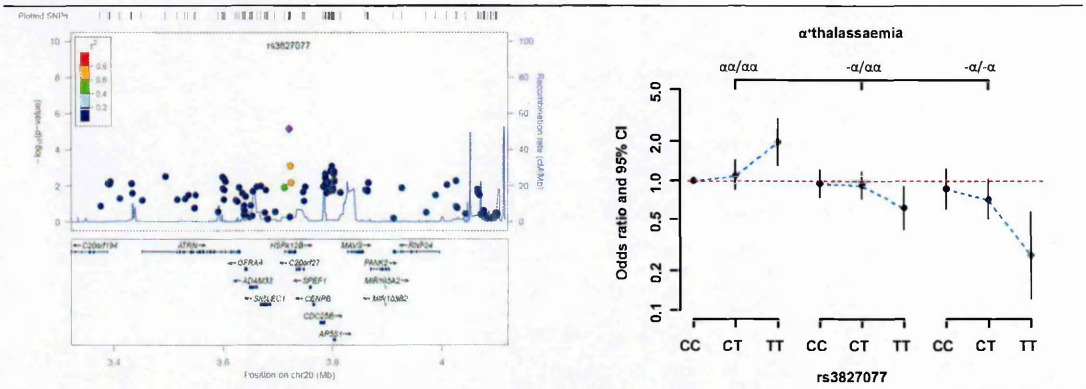


The left panel shows +/- 400kb around the top hit SNP for the α^+thalassaemia GWIS, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio ±95%CI (y-axis). The red line shows the point of no effect.

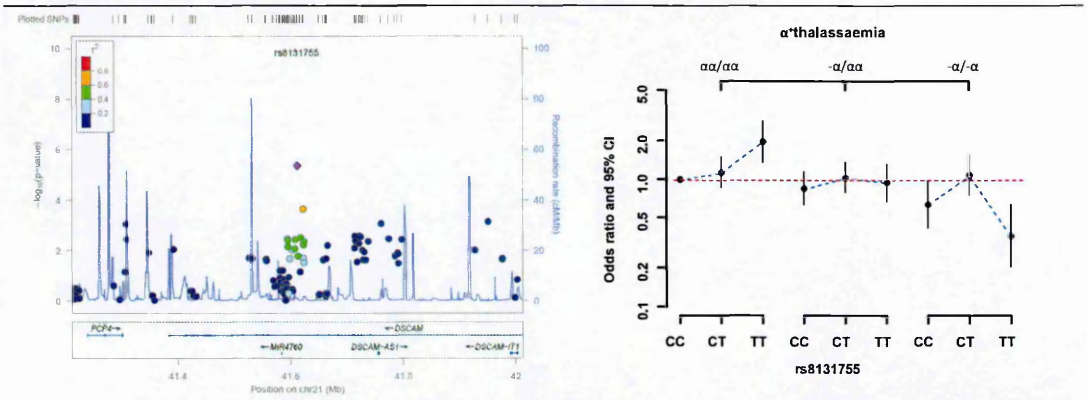
Region 7. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs8060012 on chromosome 16.



Region 8. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs3827077 on chromosome 20.



Region 9. A regional plot and a Forrest plot of interaction between α^+thalassaemia and rs8131755 on chromosome 21.



The left panel shows +/- 400kb around the top hit SNP for the α^+thalassaemia GWIS, which had the smallest p-value in the region. SNPs are coloured based on their LD (r^2), with the the top hit SNP. The –log10 P values for the SNPs are shown in the upper part of the plot. The bottom section of the plot shows the fine scale recombination rates, and genes are marked by horizontal blue arrows and arrowheads. The right panel is a Forrest plot showing interaction pattern between HbS and the top hit SNP. For each combination of genotypes (x-axis), I computed the odds ratio ±95%CI (y-axis). The red line shows the point of no effect.

References

[1] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay, "The global distribution of clinical episodes of Plasmodium falciparum malaria," *Nature,* vol. 434, pp. 214-7, Mar 10 2005.

[2] World Health Organization, "World Malaria Report 2014," World Health Organization, Geneva, 2014.

[3] World Health Organization, "World Malaria Report 2010," World Health Organization, Geneva, 2010.

[4] P. W. Gething, A. P. Patil, D. L. Smith, C. A. Guerra, I. R. Elyazar, G. L. Johnston, A. J. Tatem, and S. I. Hay, "A new world malaria map: Plasmodium falciparum endemicity in 2010," *Malar J,* vol. 10, p. 378, 2011.

[5] Kenya National Bureau of Statistics, "Kenya Population and Housing Census 2009," 2010.

[6] World Health Organization, "WHO Country Health System Fact Sheet," Geneva, 2006.

[7] Office for National statistics, "Child Mortality Statistics: Childhood, Infant and Perinatal,2012," Office for National statistics, United Kingdom, 2014.

[8] Division of malaria control, "Epidemiology of Malaria in Kenya " Division of malaria control, Nairobi, 2014.

[9] Malaria Genomic Epidemiological, Network, "A global network for investigating the genomic epidemiology of malaria," *Nature,* vol. 456, pp. 732-7, Dec 11 2008.

[10] President's Malaria Initiative, "Malaria Operational Plan (MOP)," Nairobi, 2009.

[11] Kenya National Bureau of Statistics and ICF Macro, "Kenya Malaria Indicator Survey," Nairobi, 2010.

[12] B. M. Greenwood, K. Bojang, C. J. Whitty, and G. A. Targett, "Malaria," *Lancet,* vol. 365, pp. 1487-98, Apr 23-29 2005.

[13] T. Bousema and C. Drakeley, "Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination," *Clin Microbiol Rev,* vol. 24, pp. 377-410, Apr 2011.

[14] B. Greenwood and T. Mutabingwa, "Malaria in 2002," *Nature,* vol. 415, pp. 670-2, Feb 7 2002.

[15] World Health Organization, "WHO releases new malaria guidelines for treatment and procurement of medicines," Geneva, 2010.

[16] "Severe malaria," *Trop Med Int Health,* vol. 19 Suppl 1, pp. 7-131, Sep 2014.

[17] A. C. Allison, "The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria," *Trans R Soc Trop Med Hyg,* vol. 48, pp. 312-8, Jul 1954.

[18] R. W. Snow, I. Bastos de Azevedo, B. S. Lowe, E. W. Kabiru, C. G. Nevill, S. Mwankusye, G. Kassiga, K. Marsh, and T. Teuscher, "Severe childhood malaria in two areas of markedly different falciparum transmission in east Africa," *Acta Trop,* vol. 57, pp. 289-300, Sep 1994.

[19] J. F. Trape, A. Zoulani, and M. C. Quinet, "Assessment of the incidence and prevalence of clinical malaria in semi-immune children exposed to intense and perennial transmission," *Am J Epidemiol,* vol. 126, pp. 193-201, Aug 1987.

[20] B. Greenwood, K. Marsh, and R. Snow, "Why do some African children develop severe malaria?," *Parasitol Today,* vol. 7, pp. 277-81, Oct 1991.

[21] C. Menendez, A. F. Fleming, and P. L. Alonso, "Malaria-related anaemia," *Parasitol Today,* vol. 16, pp. 469-76, Nov 2000.

[22] K. Marsh, D. Forster, C. Waruiru, I. Mwangi, M. Winstanley, V. Marsh, C. Newton, P. Winstanley, P. Warn, N. Peshu, and et al., "Indicators of life-threatening malaria in African children," *N Engl J Med,* vol. 332, pp. 1399-404, May 25 1995.

[23] J. C. Calis, K. S. Phiri, E. B. Faragher, B. J. Brabin, I. Bates, L. E. Cuevas, R. J. de Haan, A. I. Phiri, P. Malange, M. Khoka, P. J. Hulshof, L. van Lieshout, M. G. Beld, Y. Y. Teo, K. A. Rockett, A. Richardson, *et.al,* "Severe anemia in Malawian children," *N Engl J Med,* vol. 358, pp. 888-99, Feb 28 2008.

[24] B. M. Greenwood, "The epidemiology of malaria," *Ann Trop Med Parasitol,* vol. 91, pp. 763-9, Oct 1997.

[25] R. E. Phillips and T. Solomon, "Cerebral malaria in children," *Lancet,* vol. 336, pp. 1355-60, Dec 1 1990.

[26] C. R. Newton, T. E. Taylor, and R. O. Whitten, "Pathophysiology of fatal falciparum malaria in African children," *Am J Trop Med Hyg,* vol. 58, pp. 673-83, May 1998.

[27] K. Maitland and K. Marsh, "Pathophysiology of severe malaria in children," *Acta Trop,* vol. 90, pp. 131-40, Apr 2004.

[28] J. M. Olson, J. S. Witte, and R. C. Elston, "Genetic mapping of complex traits," *Stat Med,* vol. 18, pp. 2961-81, Nov 15 1999.

[29] Coriell Institute for medical research, *DNA, Genes and SNPs.* Available: https://www.coriell.org/personalized-medicine/dna-genes-and-snps

[30] P. Schlagenhauf, "Malaria: from prehistory to present," *Infect Dis Clin North Am,* vol. 18, pp. 189-205, table of contents, Jun 2004.

[31] D. J. Weatherall and J. B. Clegg, "Genetic variability in response to infection: malaria and after," *Genes Immun,* vol. 3, pp. 331-7, Sep 2002.

[32] A. Jepson, F. Sisay-Joof, W. Banya, M. Hassan-King, A. Frodsham, S. Bennett, A. V. Hill, and H. Whittle, "Genetic linkage of mild malaria to the major histocompatibility complex in Gambian children: study of affected sibling pairs," *BMJ,* vol. 315, pp. 96-7, Jul 12 1997.

[33] M. J. Mackinnon, D. M. Gunawardena, J. Rajakaruna, S. Weerasingha, K. N. Mendis, and R. Carter, "Quantifying genetic and nongenetic contributions to malarial infection in a Sri Lankan population," *Proc Natl Acad Sci U S A,* vol. 97, pp. 12661-6, Nov 7 2000.

[34] M. J. Mackinnon, T. W. Mwangi, R. W. Snow, K. Marsh, and T. N. Williams, "Heritability of malaria in Africa," *PLoS Med,* vol. 2, p. e340, Dec 2005.

[35] W.Phimpraphi, R. Paul, B. Witoonpanich, C. Turbpaiboon, C. Peerapittayamongkol, C. Louicharoen, I. Casademont, S. Tungpradabkul, S. Krudsood, J. Kaewkunwal, T. Sura, S. Looareesuwan, P. Singhasivanon, and A. Sakuntabhai, "Heritability of P. falciparum and P. vivax malaria in a Karen population in Thailand," *PLoS One,* vol. 3, p. e3887, 2008.

[36] T. N. Williams, T. W. Mwangi, S. Wambua, N. D. Alexander, M. Kortok, R. W. Snow, and K. Marsh, "Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases," *J Infect Dis,* vol. 192, pp. 178-86, Jul 1 2005.

[37] R. Hutagalung, P. Wilairatana, S. Looareesuwan, G. M. Brittenham, M. Aikawa, and V. R. Gordeuk, "Influence of hemoglobin E trait on the severity of Falciparum malaria," *J Infect Dis,* vol. 179, pp. 283-6, Jan 1999.

[38] A. Agarwal, A. Guindo, Y. Cissoko, J. G. Taylor, D. Coulibaly, A. Kone, K. Kayentao, A. Djimde, C. V. Plowe, O. Doumbo, T. E. Wellems, and D. Diallo, "Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S," *Blood,* vol. 96, pp. 2358-63, Oct 1 2000.

[39] D. J. Weatherall, "Thalassaemia and malaria, revisited," *Ann Trop Med Parasitol,* vol. 91, pp. 885-90, Oct 1997.

[40] F. P. Mockenhaupt, S. Ehrhardt, S. Gellert, R. N. Otchwemah, E. Dietz, S. D. Anemana, and U. Bienzle, "Alpha(+)-thalassemia protects African children from severe malaria," *Blood,* vol. 104, pp. 2003-6, Oct 1 2004.

[41] T. N. Williams, S. Wambua, S. Uyoga, A. Macharia, J. K. Mwacharo, C. R. Newton, and K. Maitland, "Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya," *Blood,* vol. 106, pp. 368-71, Jul 1 2005.

[42] T. G. Clark, A. E. Fry, S. Auburn, S. Campino, M. Diakite, A. Green, A. Richardson, Y. Y. Teo, K. Small, J. Wilson, M. Jallow, F. Sisay-Joof, M. Pinder, P. Sabeti, D. P. Kwiatkowski, and K. A. Rockett, "Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility," *Eur J Hum Genet,* vol. 17, pp. 1080-5, Aug 2009.

[43] A. Guindo, R. M. Fairhurst, O. K. Doumbo, T. E. Wellems, and D. A. Diallo, "X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria," *PLoS Med,* vol. 4, p. e66, Mar 2007.

[44] L. C. Foo, V. Rekhraj, G. L. Chiang, and J. W. Mak, "Ovalocytosis protects against severe malaria parasitemia in the Malayan aborigines," *Am J Trop Med Hyg,* vol. 47, pp. 271-5, Sep 1992.

[45] N. Malaria Genomic Epidemiology and N. Malaria Genomic Epidemiology, "Reappraisal of known malaria resistance loci in a large multicenter study," *Nat Genet,* vol. 46, pp. 1197-204, Nov 2014.

[46] A. V. Hill, C. E. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, A. J. McMichael, and B. M. Greenwood, "Common west African HLA antigens are associated with protection from severe malaria," *Nature,* vol. 352, pp. 595-600, Aug 15 1991.

[47] W. McGuire, A. V. Hill, C. E. Allsopp, B. M. Greenwood, and D. Kwiatkowski, "Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria," *Nature,* vol. 371, pp. 508-10, Oct 6 1994.

[48] A. G. Wilson, J. A. Symons, T. L. McDowell, H. O. McDevitt, and G. W. Duff, "Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation," *Proc Natl Acad Sci U S A,* vol. 94, pp. 3195-9, Apr 1 1997.

[49] J. C. Knight, I. Udalova, A. V. Hill, B. M. Greenwood, N. Peshu, K. Marsh, and D. Kwiatkowski, "A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria," *Nat Genet,* vol. 22, pp. 145-50, Jun 1999.

[50] G. Morahan, C. S. Boutlis, D. Huang, A. Pain, J. R. Saunders, M. R. Hobbs, D. L. Granger, J. B. Weinberg, N. Peshu, E. D. Mwaikambo, K. Marsh, D. J. Roberts, and N. M. Anstey, "A promoter polymorphism in the gene encoding interleukin-12 p40 (IL12B) is associated with mortality from cerebral malaria and with reduced nitric oxide production," *Genes Immun,* vol. 3, pp. 414-8, Nov 2002.

[51] C. Aucan, A. J. Walley, B. J. Hennig, J. Fitness, A. Frodsham, L. Zhang, D. Kwiatkowski, and A. V. Hill, "Interferon-alpha receptor-1 (IFNAR1) variants are associated with protection against cerebral malaria in the Gambia," *Genes Immun,* vol. 4, pp. 275-82, Jun 2003.

[52] B. A. Gyan, B. Goka, J. T. Cvetkovic, J. L. Kurtzhals, V. Adabayeri, H. Perlmann, A. K. Lefvert, B. D. Akanmori, and M. Troye-Blomberg, "Allelic polymorphisms in the repeat and promoter regions of the interleukin-4 gene and malaria severity in Ghanaian children," *Clin Exp Immunol,* vol. 138, pp. 145-50, Oct 2004.

[53] O. K. Amodu, A. A. Adeyemo, O. O. Ayoola, R. A. Gbadegesin, A. E. Orimadegun, A. K. Akinsola, P. E. Olumese, and O. O. Omotade, "Genetic diversity of the msp-1 locus and symptomatic malaria in south-west Nigeria," *Acta Trop,* vol. 95, pp. 226-32, Sep 2005.

[54] A. E. Fry, A. Ghansa, K. S. Small, A. Palma, S. Auburn, M. Diakite, A. Green, S. Campino, Y. Y. Teo, T. G. Clark, A. E. Jeffreys, J. Wilson, M. Jallow, F. Sisay-Joof, M. Pinder, M. J. Griffiths, N. Peshu, T. N. Williams *et. al,* "Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes," *Hum Mol Genet,* vol. 18, pp. 2683-92, Jul 15 2009.

[55] A. Sakuntabhai, R. Ndiaye, I. Casademont, C. Peerapittayamongkol, C. Rogier, P. Tortevoye, A. Tall, R. Paul, C. Turbpaiboon, W. Phimpraphi, J. F. Trape, A. Spiegel, S. Heath, O. Mercereau-Puijalon, A. Dieye, and C. Julier, "Genetic determination and linkage mapping of Plasmodium falciparum malaria related traits in Senegal," *PLoS One,* vol. 3, p. e2000, 2008.

[56] L. Flori, B. Kumulungui, C. Aucan, C. Esnault, A. S. Traore, F. Fumoux, and P. Rihet, "Linkage and association between Plasmodium falciparum blood infection levels and chromosome 5q31-q33," *Genes Immun,* vol. 4, pp. 265-8, Jun 2003.

[57] P. D. Crompton, B. Traore, K. Kayentao, S. Doumbo, A. Ongoiba, S. A. Diakite, M. A. Krause, D. Doumtabe, Y. Kone, G. Weiss, C. Y. Huang, S. Doumbia, A. Guindo, R. M. Fairhurst, L. H. Miller, S. K. Pierce, and O. K. Doumbo, "Sickle cell trait is associated with a delayed onset of malaria: implications for time-to-event analysis in clinical studies of malaria," *J Infect Dis,* vol. 198, pp. 1265-75, Nov 1 2008.

[58] A. Garcia, S. Marquet, B. Bucheton, D. Hillaire, M. Cot, N. Fievet, A. J. Dessein, and L. Abel, "Linkage analysis of blood Plasmodium falciparum levels: interest of the 5q31-q33 chromosome region," *Am J Trop Med Hyg,* vol. 58, pp. 705-9, Jun 1998.

[59] P. Rihet, Y. Traore, L. Abel, C. Aucan, T. Traore-Leroux, and F. Fumoux, "Malaria in humans: Plasmodium falciparum blood infection levels are linked to chromosome 5q31-q33," *Am J Hum Genet,* vol. 63, pp. 498-505, Aug 1998.

[60] D. P. Eisen, L. Wang, H. Jouin, E. E. Murhandarwati, C. G. Black, O. Mercereau-Puijalon, and R. L. Coppel, "Antibodies elicited in adults by a primary Plasmodium falciparum blood-stage infection recognize different epitopes compared with immune individuals," *Malar J,* vol. 6, p. 86, 2007.

[61] C. C. Khor, F. O. Vannberg, S. J. Chapman, A. Walley, C. Aucan, H. Loke, N. J. White, T. Peto, L. K. Khor, D. Kwiatkowski, N. Day, A. Scott, J. A. Berkley, K. Marsh, N. Peshu, K. Maitland, T. N. Williams, and A. V. Hill, "Positive replication and linkage disequilibrium mapping of the chromosome 21q22.1 malaria susceptibility locus," *Genes Immun,* vol. 8, pp. 570-6, Oct 2007.

[62] J. May, J. A. Evans, C. Timmann, C. Ehmen, W. Busch, T. Thye, T. Agbenyega, and R. D. Horstmann, "Hemoglobin variants and disease manifestations in severe falciparum malaria," *JAMA,* vol. 297, pp. 2220-6, May 23 2007.

[63] V. Thathy, J. M. Moulds, B. Guyah, W. Otieno, and J. A. Stoute, "Complement receptor 1 polymorphisms associated with resistance to severe malaria in Kenya," *Malar J,* vol. 4, p. 54, 2005.

[64] R. Bellamy, D. Kwiatkowski, and A. V. Hill, "Absence of an association between intercellular adhesion molecule 1, complement receptor 1 and interleukin 1 receptor antagonist gene polymorphisms and severe malaria in a West African population," *Trans R Soc Trop Med Hyg,* vol. 92, pp. 312-6, May-Jun 1998.

[65] J. A. Rowe, A. Raza, D. A. Diallo, M. Baby, B. Poudiougo, D. Coulibaly, I. A. Cockburn, J. Middleton, K. E. Lyke, C. V. Plowe, O. K. Doumbo, and J. M. Moulds, "Erythrocyte CR1 expression level does not correlate with a HindIII restriction fragment length polymorphism in Africans; implications for studies on malaria susceptibility," *Genes Immun,* vol. 3, pp. 497-500, Dec 2002.

[66] S. Wambua, T. W. Mwangi, M. Kortok, S. M. Uyoga, A. W. Macharia, J. K. Mwacharo, D. J. Weatherall, R. W. Snow, K. Marsh, and T. N. Williams, "The effect of alpha+-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya," *PLoS Med,* vol. 3, p. e158, May 2006.

[67] J. A. Rowe, I. G. Handel, M. A. Thera, A. M. Deans, K. E. Lyke, A. Kone, D. A. Diallo, A. Raza, O. Kai, K. Marsh, C. V. Plowe, O. K. Doumbo, and J. M. Moulds, "Blood group O protects against severe Plasmodium falciparum malaria through the mechanism of reduced rosetting," *Proc Natl Acad Sci U S A,* vol. 104, pp. 17471-6, Oct 30 2007.

[68] P. A. Zimmerman, J. Fitness, J. M. Moulds, D. T. McNamara, L. J. Kasehagen, J. A. Rowe, and A. V. Hill, "CR1 Knops blood group alleles are not associated with severe malaria in the Gambia," *Genes Immun,* vol. 4, pp. 368-73, Jul 2003.

[69] J. R. Ryan, J. A. Stoute, J. Amon, R. F. Dunton, R. Mtalib, J. Koros, B. Owour, S. Luckhart, R. A. Wirtz, J. W. Barnwell, and R. Rosenberg, "Evidence for transmission of Plasmodium vivax among a duffy antigen negative population in Western Kenya," *Am J Trop Med Hyg,* vol. 75, pp. 575-81, Oct 2006.

[70] G. S. Cooke, C. Aucan, A. J. Walley, S. Segal, B. M. Greenwood, D. P. Kwiatkowski, and A. V. Hill, "Association of Fcgamma receptor IIa (CD32) polymorphism with severe malaria in West Africa," *Am J Trop Med Hyg,* vol. 69, pp. 565-8, Dec 2003.

[71] A. Pain, B. C. Urban, O. Kai, C. Casals-Pascual, J. Shafi, K. Marsh, and D. J. Roberts, "A non-sense mutation in Cd36 gene is associated with protection from severe malaria," *Lancet,* vol. 357, pp. 1502-3, May 12 2001.

[72] P. Sabeti, S. Usen, S. Farhadian, M. Jallow, T. Doherty, M. Newport, M. Pinder, R. Ward, and D. Kwiatkowski, "CD40L association with protection from severe malaria," *Genes Immun,* vol. 3, pp. 286-91, Aug 2002.

[73] M. Jallow, Y. Y. Teo, K. S. Small, K. A. Rockett, P. Deloukas, T. G. Clark, K. Kivinen, K. A. Bojang, D. J. Conway, M. Pinder, G. Sirugo, F. Sisay-Joof, S. Usen, S. Auburn, S. J. Bumpstead, S. Campino, A. Coffey, A. Dunham, A. *et al,* "Genome-wide and fine-resolution association analysis of malaria in West Africa," *Nat Genet,* vol. 41, pp. 657-65, Jun 2009.

[74] A. Nasr, N. C. Iriemenam, M. Troye-Blomberg, H. A. Giha, H. A. Balogun, O. F. Osman, S. M. Montgomery, G. ElGhazali, and K. Berzins, "Fc gamma receptor IIa (CD32) polymorphism and antibody responses to asexual blood-stage antigens of Plasmodium falciparum malaria in Sudanese patients," *Scand J Immunol,* vol. 66, pp. 87-96, Jul 2007.

[75] C. Ouma, C. C. Keller, D. A. Opondo, T. Were, R. O. Otieno, M. F. Otieno, A. S. Orago, J. M. Ong'Echa, J. M. Vulule, R. E. Ferrell, and D. J. Perkins, "Association of FCgamma receptor IIA (CD32) polymorphism with malarial anemia and high-density parasitemia in infants and young children," *Am J Trop Med Hyg,* vol. 74, pp. 573-7, Apr 2006.

[76] K. Schuldt, C. Esser, J. Evans, J. May, C. Timmann, C. Ehmen, W. Loag, D. Ansong, A. Ziegler, T. Agbenyega, C. G. Meyer, and R. D. Horstmann, "FCGR2A functional genetic variant associated with susceptibility to severe malarial anaemia in Ghanaian children," *J Med Genet,* vol. 47, pp. 471-5, Jul 2010.

[77] Y. P. Shi, B. L. Nahlen, S. Kariuki, K. B. Urdahl, P. D. McElroy, J. M. Roberts, and A. A. Lal, "Fcgamma receptor IIa (CD32) polymorphism is associated with protection of infants against high-density Plasmodium falciparum infection. VII. Asembo Bay Cohort Project," *J Infect Dis,* vol. 184, pp. 107-11, Jul 1 2001.

[78] G. Ayodo, A. L. Price, A. Keinan, A. Ajwang, M. F. Otieno, A. S. Orago, N. Patterson, and D. Reich, "Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants," *Am J Hum Genet,* vol. 81, pp. 234-42, Aug 2007.

[79] A. Muehlenbachs, M. Fried, J. Lachowitzer, T. K. Mutabingwa, and P. E. Duffy, "Natural selection of FLT1 alleles and their association with malaria resistance in utero," *Proc Natl Acad Sci U S A,* vol. 105, pp. 14488-91, Sep 23 2008.

[80] M. Sikora, A. Ferrer-Admetlla, H. Laayouni, C. Menendez, A. Mayor, A. Bardaji, B. Sigauque, I. Mandomando, P. L. Alonso, J. Bertranpetit, and F. Casals, "A variant in the gene FUT9 is associated with susceptibility to placental malaria infection," *Hum Mol Genet,* vol. 18, pp. 3136-44, Aug 15 2009.

[81] T. D. Clark, B. Greenhouse, D. Njama-Meya, B. Nzarubara, C. Maiteki-Sebuguzi, S. G. Staedke, E. Seto, M. R. Kamya, P. J. Rosenthal, and G. Dorsey, "Factors determining the heterogeneity of malaria incidence in children in Kampala, Uganda," *J Infect Dis,* vol. 198, pp. 393-400, Aug 1 2008.

[82] M. K. Johnson, T. D. Clark, D. Njama-Meya, P. J. Rosenthal, and S. Parikh, "Impact of the method of G6PD deficiency assessment on genetic association studies of malaria susceptibility," *PLoS One,* vol. 4, p. e7246, 2009.

[83] F. Migot-Nabias, L. E. Mombo, A. J. Luty, B. Dubois, R. Nabias, C. Bisseye, P. Millet, C. Y. Lu, and P. Deloron, "Human genetic factors related to susceptibility to mild malaria in Gabon," *Genes Immun,* vol. 1, pp. 435-41, Oct 2000.

[84] L. E. Mombo, F. Ntoumi, C. Bisseye, S. Ossari, C. Y. Lu, R. L. Nagel, and R. Krishnamoorthy, "Human genetic polymorphisms and asymptomatic Plasmodium falciparum malaria in Gabonese schoolchildren," *Am J Trop Med Hyg,* vol. 68, pp. 186-90, Feb 2003.

[85] S. Parikh, G. Dorsey, and P. J. Rosenthal, "Host polymorphisms and the incidence of malaria in Ugandan children," *Am J Trop Med Hyg,* vol. 71, pp. 750-3, Dec 2004.

[86] S. Shekalaghe, M. Alifrangis, C. Mwanziva, A. Enevold, S. Mwakalinga, H. Mkali, R. Kavishe, A. Manjurano, R. Sauerwein, C. Drakeley, and T. Bousema, "Low density parasitaemia, red blood cell polymorphisms and Plasmodium falciparum specific immune responses in a low endemic area in northern Tanzania," *BMC Infect Dis,* vol. 9, p. 69, 2009.

[87] M. Vafa, M. Troye-Blomberg, J. Anchang, A. Garcia, and F. Migot-Nabias, "Multiplicity of Plasmodium falciparum infection in asymptomatic children in Senegal: relation to transmission, age and erythrocyte variants," *Malar J,* vol. 7, p. 17, 2008.

[88] B. Lell, J. May, R. J. Schmidt-Ott, L. G. Lehman, D. Luckner, B. Greve, P. Matousek, D. Schmid, K. Herbich, F. P. Mockenhaupt, C. G. Meyer, U. Bienzle, and P. G. Kremsner, "The role of red blood cell polymorphisms in resistance and susceptibility to malaria," *Clin Infect Dis,* vol. 28, pp. 794-9, Apr 1999.

[89] R. A. Kavishe, J. B. Koenderink, M. B. McCall, W. H. Peters, B. Mulder, C. C. Hermsen, R. W. Sauerwein, F. G. Russel, and A. J. Van der Ven, "Short report: Severe Plasmodium falciparum malaria in Cameroon: associated with the glutathione S-transferase M1 null genotype," *Am J Trop Med Hyg,* vol. 75, pp. 827-9, Nov 2006.

[90] R. A. Kavishe, T. Bousema, S. A. Shekalaghe, R. W. Sauerwein, F. W. Mosha, A. J. van der Ven, F. G. Russel, and J. B. Koenderink, "Common genotypic polymorphisms in glutathione S-transferases in mild and severe falciparum malaria in Tanzanian children," *Am J Trop Med Hyg,* vol. 81, pp. 363-5, Aug 2009.

[91] J. Baum, M. Pinder, and D. J. Conway, "Erythrocyte invasion phenotypes of Plasmodium falciparum in The Gambia," *Infect Immun,* vol. 71, pp. 1856-63, Apr 2003.

[92] S. Auburn, M. Diakite, A. E. Fry, A. Ghansah, S. Campino, A. Richardson, M. Jallow, F. Sisay-Joof, M. Pinder, M. J. Griffiths, N. Peshu, T. N. Williams, K. Marsh, M. E. Molyneux, T. E. Taylor, K. A. Koram, A. R. Oduro, W. O. Rogers, K. A. Rockett, K. Haldar, and D. P. Kwiatkowski, "Association of the GNAS locus with severe malaria," *Hum Genet,* vol. 124, pp. 499-506, Dec 2008.

[93] S. H. Atkinson, T. W. Mwangi, S. M. Uyoga, E. Ogada, A. W. Macharia, K. Marsh, A. M. Prentice, and T. N. Williams, "The haptoglobin 2-2 genotype is associated with a reduced incidence of Plasmodium falciparum malaria in children on the coast of Kenya," *Clin Infect Dis,* vol. 44, pp. 802-9, Mar 15 2007.

[94] S. E. Cox, C. P. Doherty, S. H. Atkinson, C. V. Nweneka, A. J. Fulford, G. Sirugo, K. A. Rockett, D. P. Kwiatkowski, and A. M. Prentice, "Haptoglobin genotype, anaemia and malaria in Gambian children," *Trop Med Int Health,* vol. 13, pp. 76-82, Jan 2008.

[95] J. T. Minang, B. A. Gyan, J. K. Anchang, M. Troye-Blomberg, H. Perlmann, and E. A. Achidi, "Haptoglobin phenotypes and malaria infection in pregnant women at delivery in western Cameroon," *Acta Trop,* vol. 90, pp. 107-14, Mar 2004.

[96] I. K. Quaye, F. A. Ekuban, B. Q. Goka, V. Adabayeri, J. A. Kurtzhals, B. Gyan, N. A. Ankrah, L. Hviid, and B. D. Akanmori, "Haptoglobin 1-1 is associated with susceptibility to severe Plasmodium falciparum malaria," *Trans R Soc Trop Med Hyg,* vol. 94, pp. 216-9, Mar-Apr 2000.

[97] C. Aucan, A. J. Walley, and A. V. Hill, "Common apolipoprotein E polymorphisms and risk of clinical malaria in the Gambia," *J Med Genet,* vol. 41, pp. 21-4, Jan 2004.

[98] U. Bienzle, O. Ayeni, A. O. Lucas, and L. Luzzatto, "Glucose-6-phosphate dehydrogenase and malaria. Greater resistance of females heterozygous for enzyme deficiency and of males with non-deficient variant," *Lancet,* vol. 1, pp. 107-10, Jan 15 1972.

[99] M. Aidoo, D. J. Terlouw, M. S. Kolczak, P. D. McElroy, F. O. ter Kuile, S. Kariuki, B. L. Nahlen, A. A. Lal, and V. Udhayakumar, "Protective effects of the sickle cell gene against malaria morbidity and mortality," *Lancet,* vol. 359, pp. 1311-2, Apr 13 2002.

[100] A. N. Komba, J. Makani, M. Sadarangani, T. Ajala-Agbo, J. A. Berkley, C. R. Newton, K. Marsh, and T. N. Williams, "Malaria as a cause of morbidity and mortality in children with homozygous sickle cell disease on the coast of Kenya," *Clin Infect Dis,* vol. 49, pp. 216-22, Jul 15 2009.

[101] B. Kreuels, S. Ehrhardt, C. Kreuzberg, S. Adjei, R. Kobbe, G. D. Burchard, C. Ehmen, M. Ayim, O. Adjei, and J. May, "Sickle cell trait (HbAS) and stunting in children below two years of age in an area of high malaria transmission," *Malar J,* vol. 8, p. 16, 2009.

[102] V. D. Mangano, G. Luoni, K. A. Rockett, B. S. Sirima, A. Konate, J. Forton, T. G. Clark, G. Bancone, E. Sadighi Akha, D. P. Kwiatkowski, and D. Modiano, "Interferon regulatory factor-1 polymorphisms are associated with the control of Plasmodium falciparum infection," *Genes Immun,* vol. 9, pp. 122-9, Mar 2008.

[103] T. N. Williams, T. W. Mwangi, S. Wambua, T. E. Peto, D. J. Weatherall, S. Gupta, M. Recker, B. S. Penman, S. Uyoga, A. Macharia, J. K. Mwacharo, R. W. Snow, and K. Marsh, "Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait," *Nat Genet,* vol. 37, pp. 1253-7, Nov 2005.

[104] D. Fernandez-Reyes, A. G. Craig, S. A. Kyes, N. Peshu, R. W. Snow, A. R. Berendt, K. Marsh, and C. I. Newbold, "A high frequency African coding polymorphism in the N-terminal domain of ICAM-1 predisposing to cerebral malaria in Kenya," *Hum Mol Genet,* vol. 6, pp. 1357-60, Aug 1997.

[105] J. F. Kun, J. Klabunde, B. Lell, D. Luckner, M. Alpers, J. May, C. Meyer, and P. G. Kremsner, "Association of the ICAM-1Kilifi mutation with protection

against severe malaria in Lambarene, Gabon," *Am J Trop Med Hyg,* vol. 61, pp. 776-9, Nov 1999.

[106] A. E. Fry, S. Auburn, M. Diakite, A. Green, A. Richardson, J. Wilson, M. Jallow, F. Sisay-Joof, M. Pinder, M. J. Griffiths, N. Peshu, T. N. Williams, K. Marsh, M. E. Molyneux, T. E. Taylor, K. A. Rockett, and D. P. Kwiatkowski, "Variation in the ICAM1 gene is not associated with severe malaria phenotypes," *Genes Immun,* vol. 9, pp. 462-9, Jul 2008.

[107] M. Diakite, T. G. Clark, S. Auburn, S. Campino, A. E. Fry, A. Green, A. P. Morris, A. Richardson, M. Jallow, F. Sisay-Joof, M. Pinder, D. P. Kwiatkowski, and K. A. Rockett, "A genetic association study in the Gambia using tagging polymorphisms in the major histocompatibility complex class III region implicates a HLA-B associated transcript 2 polymorphism in severe malaria susceptibility," *Hum Genet,* vol. 125, pp. 105-9, Feb 2009.

[108] C. Tena-Tomas, I. de Messias-Reason, H. Song le, J. Tomiuk, P. G. Kemsner, and J. F. Kun, "A globally occurring indel polymorphism in the promoter of the IFNA2 gene is not associated with severity of malaria but with the positivity rate of HCV," *BMC Genet,* vol. 9, p. 80, 2008.

[109] S. Cabantous, B. Poudiougou, A. Traore, M. Keita, M. B. Cisse, O. Doumbo, A. J. Dessein, and S. Marquet, "Evidence that interferon-gamma plays a protective role during cerebral malaria," *J Infect Dis,* vol. 192, pp. 854-60, Sep 1 2005.

[110] O. Koch, A. Awomoyi, S. Usen, M. Jallow, A. Richardson, J. Hull, M. Pinder, M. Newport, and D. Kwiatkowski, "IFNGR1 gene promoter polymorphisms and susceptibility to cerebral malaria," *J Infect Dis,* vol. 185, pp. 1684-7, Jun 1 2002.

[111] A. J. Walley, C. Aucan, D. Kwiatkowski, and A. V. Hill, "Interleukin-1 gene cluster polymorphisms and susceptibility to clinical malaria in a Gambian case-control study," *Eur J Hum Genet,* vol. 12, pp. 132-8, Feb 2004.

[112] B. Gyan, B. Goka, J. T. Cvetkovic, H. Perlmann, A. K. Lefvert, B. Akanmori, and M. Troye-Blomberg, "Polymorphisms in interleukin-1beta and interleukin-1 receptor antagonist genes and malaria in Ghanaian children," *Scand J Immunol,* vol. 56, pp. 619-22, Dec 2002.

[113] C. Ouma, G. C. Davenport, G. A. Awandare, C. C. Keller, T. Were, M. F. Otieno, J. M. Vulule, J. Martinson, J. M. Ong'echa, R. E. Ferrell, and D. J. Perkins, "Polymorphic variability in the interleukin (IL)-1beta promoter conditions susceptibility to severe malarial anemia and functional changes in IL-1beta production," *J Infect Dis,* vol. 198, pp. 1219-26, Oct 15 2008.

[114] D. Carpenter, I. Rooth, A. Farnert, H. Abushama, R. J. Quinnell, and M. A. Shaw, "Genetics of susceptibility to malaria related phenotypes," *Infect Genet Evol,* vol. 9, pp. 97-103, Jan 2009.

[115] M. Vafa, B. Maiga, K. Berzins, M. Hayano, S. Bereczky, A. Dolo, M. Daou, C. Arama, B. Kouriba, A. Farnert, O. K. Doumbo, and M. Troye-Blomberg, "Associations between the IL-4 -590 T allele and Plasmodium falciparum infection prevalence in asymptomatic Fulani of Mali," *Microbes Infect,* vol. 9, pp. 1043-8, Jul 2007.

[116] F. Verra, V. D. Mangano, and D. Modiano, "Genetics of susceptibility to Plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies," *Parasite Immunol,* vol. 31, pp. 234-53, May 2009.

[117] J. N. Wilson, K. Rockett, M. Jallow, M. Pinder, F. Sisay-Joof, M. Newport, J. Newton, and D. Kwiatkowski, "Analysis of IL10 haplotypic associations with severe malaria," *Genes Immun,* vol. 6, pp. 462-6, Sep 2005.

[118] S. Marquet, O. Doumbo, S. Cabantous, B. Poudiougou, L. Argiro, I. Safeukui, S. Konate, S. Sissoko, E. Chevereau, A. Traore, M. M. Keita, C. Chevillard, L. Abel, and A. J. Dessein, "A functional promoter variant in IL12B predisposes to cerebral malaria," *Hum Mol Genet,* vol. 17, pp. 2190-5, Jul 15 2008.

[119] M. Barbier, A. Atkinson, F. Fumoux, and P. Rihet, "IL12B polymorphisms are linked but not associated with Plasmodium falciparum parasitemia: a familial study in Burkina Faso," *Genes Immun,* vol. 9, pp. 405-11, Jul 2008.

[120] O. Koch, K. Rockett, M. Jallow, M. Pinder, F. Sisay-Joof, and D. Kwiatkowski, "Investigation of malaria susceptibility determinants in the IFNG/IL26/IL22 genomic region," *Genes Immun,* vol. 6, pp. 312-8, Jun 2005.

[121] C. Timmann, J. A. Evans, I. R. Konig, A. Kleensang, F. Ruschendorf, J. Lenzen, J. Sievertsen, C. Becker, Y. Enuameh, K. O. Kwakye, E. Opoku, E. N. Browne, A. Ziegler, P. Nurnberg, and R. D. Horstmann, "Genome-wide linkage analysis of malaria infection intensity and mild disease," *PLoS Genet,* vol. 3, p. e48, Mar 23 2007.

[122] G. A. Awandare, J. J. Martinson, T. Were, C. Ouma, G. C. Davenport, J. M. Ong'echa, W. Wang, L. Leng, R. E. Ferrell, R. Bucala, and D. J. Perkins, "MIF (macrophage migration inhibitory factor) promoter polymorphisms and susceptibility to severe malarial anemia," *J Infect Dis,* vol. 200, pp. 629-37, Aug 15 2009.

[123] A. V. Hill, "Molecular epidemiology of the thalassaemias (including haemoglobin E)," *Baillieres Clin Haematol,* vol. 5, pp. 209-38, Jan 1992.

[124] A. D. Osafo-Addo, K. A. Koram, A. R. Oduro, M. Wilson, A. Hodgson, and W. O. Rogers, "HLA-DRB1*04 allele is associated with severe malaria in northern Ghana," *Am J Trop Med Hyg,* vol. 78, pp. 251-5, Feb 2008.

[125] S. Bennett, S. J. Allen, O. Olerup, D. J. Jackson, J. G. Wheeler, P. A. Rowe, E. M. Riley, and B. M. Greenwood, "Human leucocyte antigen (HLA) and malaria

morbidity in a Gambian community," *Trans R Soc Trop Med Hyg,* vol. 87, pp. 286-7, May-Jun 1993.

[126] N. F. Delahaye, M. Barbier, F. Fumoux, and P. Rihet, "Association analyses of NCR3 polymorphisms with P. falciparum mild malaria," *Microbes Infect,* vol. 9, pp. 160-6, Feb 2007.

[127] M. R. Hobbs, V. Udhayakumar, M. C. Levesque, J. Booth, J. M. Roberts, A. N. Tkachuk, A. Pole, H. Coon, S. Kariuki, B. L. Nahlen, E. D. Mwaikambo, A. L. Lal, D. L. Granger, N. M. Anstey, and J. B. Weinberg, "A new NOS2 promoter polymorphism associated with increased nitric oxide production and protection from severe malaria in Tanzanian and Kenyan children," *Lancet,* vol. 360, pp. 1468-75, Nov 9 2002.

[128] D. Burgner, S. Usen, K. Rockett, M. Jallow, H. Ackerman, A. Cervino, M. Pinder, and D. P. Kwiatkowski, "Nucleotide and haplotypic diversity of the NOS2A promoter region and its relationship to cerebral malaria," *Hum Genet,* vol. 112, pp. 379-86, Apr 2003.

[129] J. P. Cramer, F. P. Mockenhaupt, S. Ehrhardt, J. Burkhardt, R. N. Otchwemah, E. Dietz, S. Gellert, and U. Bienzle, "iNOS promoter variants and severe malaria in Ghanaian children," *Trop Med Int Health,* vol. 9, pp. 1074-80, Oct 2004.

[130] J. F. Kun, B. Mordmuller, B. Lell, L. G. Lehman, D. Luckner, and P. G. Kremsner, "Polymorphism in promoter region of inducible nitric oxide synthase gene and protection against malaria," *Lancet,* vol. 351, pp. 265-6, Jan 24 1998.

[131] M. C. Levesque, M. R. Hobbs, N. M. Anstey, T. N. Vaughn, J. A. Chancellor, A. Pole, D. J. Perkins, M. A. Misukonis, S. J. Chanock, D. L. Granger, and J. B. Weinberg, "Nitric oxide synthase type 2 promoter polymorphisms, nitric oxide production, and disease severity in Tanzanian children with malaria," *J Infect Dis,* vol. 180, pp. 1994-2002, Dec 1999.

[132] C. Casals-Pascual, S. Allen, A. Allen, O. Kai, B. Lowe, A. Pain, and D. J. Roberts, "Short report: codon 125 polymorphism of CD31 and susceptibility to malaria," *Am J Trop Med Hyg,* vol. 65, pp. 736-7, Dec 2001.

[133] V. von Kalckreuth, J. A. Evans, C. Timmann, D. Kuhn, T. Agbenyega, R. D. Horstmann, and J. May, "Promoter polymorphism of the anion-exchange protein 1 associated with severe malarial anemia and fatality," *J Infect Dis,* vol. 194, pp. 949-57, Oct 1 2006.

[134] I. J. Donaldson, J. Shefta, C. A. Lawson, J. R. Bushnell, A. W. Morgan, J. D. Isaacs, D. Carpenter, M. A. Shaw, I. Rooth, R. J. Quinnell, A. M. Zumla, W. R. Ollier, C. Z. Chintu, G. P. Muyinda, A. S. Hill, and A. W. Boylston, "Unique TCR beta-subunit variable gene haplotypes in Africans," *Immunogenetics,* vol. 53, pp. 884-93, Feb 2002.

[135] L. Hamann, O. Kumpf, R. P. Schuring, E. Alpsoy, G. Bedu-Addo, U. Bienzle, L. Oskam, F. P. Mockenhaupt, and R. R. Schumann, "Low frequency of the TIRAP S180L polymorphism in Africa, and its potential role in malaria, sepsis, and leprosy," *BMC Med Genet,* vol. 10, p. 65, 2009.

[136] S. Campino, J. Forton, S. Auburn, A. Fry, M. Diakite, A. Richardson, J. Hull, M. Jallow, F. Sisay-Joof, M. Pinder, M. E. Molyneux, T. E. Taylor, K. Rockett, T. G. Clark, and D. P. Kwiatkowski, "TLR9 polymorphisms in African populations: no association with severe malaria, but evidence of cis-variants acting on gene expression," *Malar J,* vol. 8, p. 44, 2009.

[137] L. Manco, P. Machado, D. Lopes, F. Nogueira, V. E. Do Rosario, P. L. Alonso, L. Varandas, J. Trovoada Mde, A. Amorim, and A. P. Arez, "Analysis of TPI gene promoter variation in three sub-Saharan Africa population samples," *Am J Hum Biol,* vol. 21, pp. 118-20, Jan-Feb 2009.

[138] G. Buchanan, E. Vichinsky, L. Krishnamurti, and S. Shenoy, "Severe sickle cell disease--pathophysiology and therapy," *Biol Blood Marrow Transplant,* vol. 16, pp. S64-7, Jan 2010.

[139] A. C. Allison, "Protection afforded by sickle-cell trait against subtertian malareal infection," *Br Med J,* vol. 1, pp. 290-4, Feb 6 1954.

[140] E. A. Beet, "Sickle cell disease in Northern Rhodesia," *East Afr Med J,* vol. 24, pp. 212-22, Jun 1947.

[141] R. Galanello and A. Cao, "Gene test review. Alpha-thalassemia," *Genet Med,* vol. 13, pp. 83-8, Feb 2011.

[142] P. W. Hedrick, "Selection and mutation for alpha Thalassemia in nonmalarial and malarial environments," *Ann Hum Genet,* vol. 75, pp. 468-74, Jul 2011.

[143] S. Wambua, J. Mwacharo, S. Uyoga, A. Macharia, and T. N. Williams, "Co-inheritance of alpha+-thalassaemia and sickle trait results in specific effects on haematological parameters," *Br J Haematol,* vol. 133, pp. 206-9, Apr 2006.

[144] P. Greenwell, "Blood group antigens: molecules seeking a function?," *Glycoconj J,* vol. 14, pp. 159-73, Feb 1997.

[145] C. M. Cserti and W. H. Dzik, "The ABO blood group system and Plasmodium falciparum malaria," *Blood,* vol. 110, pp. 2250-8, Oct 1 2007.

[146] L. F. C. Reid ME, *The blood group antigen factsbook.* Amsterdam ;Boston: Elsevier/Academic Press, 2003.

[147] F. Yamamoto, H. Clausen, T. White, J. Marken, and S. Hakomori, "Molecular genetic basis of the histo-blood group ABO system," *Nature,* vol. 345, pp. 229-33, May 17 1990.

[148] C. J. Uneke, "Plasmodium falciparum malaria and ABO blood group: is there any relationship?," *Parasitol Res,* vol. 100, pp. 759-65, Mar 2007.

[149] S. A. Tishkoff, R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams, and A. G. Clark, "Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance," *Science,* vol. 293, pp. 455-62, Jul 20 2001.

[150] E. Beutler, W. Kuhl, J. L. Vives-Corrons, and J. T. Prchal, "Molecular heterogeneity of glucose-6-phosphate dehydrogenase A," *Blood,* vol. 74, pp. 2550-5, Nov 15 1989.

[151] S. Piomelli, L. M. Corash, D. D. Davenport, J. Miraglia, and E. L. Amorosi, "In vivo lability of glucose-6-phosphate dehydrogenase in GdA- and GdMediterranean deficiency," *J Clin Invest,* vol. 47, pp. 940-8, Apr 1968.

[152] J. P. Ioannidis, E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis, "Replication validity of genetic association studies," *Nat Genet,* vol. 29, pp. 306-9, Nov 2001.

[153] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et. al,* International Human Genome Sequencing, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, Feb 15 2001.

[154] J. D. McPherson, M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E. R. Mardis, R. K. Wilson, *et.al,* International Human Genome Mapping, "A physical map of the human genome," *Nature,* vol. 409, pp. 934-41, Feb 15 2001.

[155] C. International HapMap, "The International HapMap Project," *Nature,* vol. 426, pp. 789-96, Dec 18 2003.

[156] C. International HapMap, "A haplotype map of the human genome," *Nature,* vol. 437, pp. 1299-320, Oct 27 2005.

[157] C. Wellcome Trust Case Control, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature,* vol. 447, pp. 661-78, Jun 7 2007.

[158] G. Band, Q. S. Le, L. Jostins, M. Pirinen, K. Kivinen, M. Jallow, F. Sisay-Joof, K. Bojang, M. Pinder, G. Sirugo *et.al,* "Imputation-based meta-analysis of severe malaria in three African populations," *PLoS Genet,* vol. 9, p. e1003509, May 2013.

[159] N. Hanchard, A. Elzein, C. Trafford, K. Rockett, M. Pinder, M. Jallow, R. Harding, D. Kwiatkowski, and C. McKenzie, "Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations," *BMC Genet,* vol. 8, p. 52, 2007.

[160] G. H. Kijak, A. M. Walsh, R. N. Koehler, N. Moqueet, L. A. Eller, M. Eller, J. R. Currier, Z. Wang, F. Wabwire-Mangen, H. N. Kibuuka, N. L. Michael, M. L. Robb, and F. E. McCutchan, "HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster," *Tissue Antigens,* vol. 73, pp. 262-9, Mar 2009.

[161] J. J. Just, M. C. King, G. Thomson, and W. Klitz, "African-American HLA class II allele and haplotype diversity," *Tissue Antigens,* vol. 48, pp. 636-44, Dec 1996.

[162] B. W., *Mendel's principles of heredity.* Cambridge: Cambridge University Press, 1909.

[163] T. F. Mackay and J. H. Moore, "Why epistasis is important for tackling complex human disease genetics," *Genome Med,* vol. 6, p. 42, 2014.

[164] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Hum Mol Genet,* vol. 11, pp. 2463-8, Oct 1 2002.

[165] J. Altmuller, L. J. Palmer, G. Fischer, H. Scherb, and M. Wjst, "Genomewide scans of complex human diseases: true linkage is hard to find," *Am J Hum Genet,* vol. 69, pp. 936-50, Nov 2001.

[166] H. J. Cordell, J. A. Todd, S. T. Bennett, Y. Kawaguchi, and M. Farrall, "Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes," *Am J Hum Genet,* vol. 57, pp. 920-34, Oct 1995.

[167] N. J. Cox, M. Frigge, D. L. Nicolae, P. Concannon, C. L. Hanis, G. I. Bell, and A. Kong, "Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans," *Nat Genet,* vol. 21, pp. 213-5, Feb 1999.

[168] J. Xu, C. D. Langefeld, S. L. Zheng, E. M. Gillanders, B. L. Chang, S. D. Isaacs, A. H. Williams, K. E. Wiley, L. Dimitrov, D. A. Meyers, P. C. Walsh, J. M. Trent, and W. B. Isaacs, "Interaction effect of PTEN and CDKN1B chromosomal regions on prostate cancer linkage," *Hum Genet,* vol. 115, pp. 255-62, Aug 2004.

[169] C. E. Aston, D. A. Ralph, D. P. Lalo, S. Manjeshwar, B. A. Gramling, D. C. DeFreese, A. D. West, D. E. Branam, L. F. Thompson, M. A. Craft, D. S. Mitchell, C. D. Shimasaki, J. J. Mulvihill, and E. R. Jupe, "Oligogenic combinations associated with breast cancer risk in women under 53 years of age," *Hum Genet,* vol. 116, pp. 208-21, Feb 2005.

[170] S. M. Williams, J. H. Addy, J. A. Phillips, 3rd, M. Dai, J. Kpodonu, J. Afful, H. Jackson, K. Joseph, F. Eason, M. M. Murray, P. Epperson, A. Aduonum, L. J.

Wong, P. A. Jose, and R. A. Felder, "Combinations of variations in multiple genes are associated with hypertension," *Hypertension,* vol. 36, pp. 2-6, Jul 2000.

[171] C. Dong, W. D. Li, D. Li, and R. A. Price, "Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21," *Eur J Hum Genet,* vol. 13, pp. 102-8, Jan 2005.

[172] J. A. Staessen, J. G. Wang, E. Brand, C. Barlassina, W. H. Birkenhager, S. M. Herrmann, R. Fagard, L. Tizzoni, and G. Bianchi, "Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population," *J Hypertens,* vol. 19, pp. 1349-58, Aug 2001.

[173] A. Balmain and C. C. Harris, "Carcinogenesis in mouse and human cells: parallels and paradoxes," *Carcinogenesis,* vol. 21, pp. 371-7, Mar 2000.

[174] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: limits of models displaying no main effect," *Am J Hum Genet,* vol. 70, pp. 461-71, Feb 2002.

[175] J. C. Pedersen and K. Berg, "Interaction between low density lipoprotein receptor (LDLR) and apolipoprotein E (apoE) alleles contributes to normal variation in lipid level," *Clin Genet,* vol. 35, pp. 331-7, May 1989.

[176] J. Xu, D. A. Meyers, C. Ober, M. N. Blumenthal, B. Mellen, K. C. Barnes, R. A. King, L. A. Lester, T. D. Howard, J. Solway, C. D. Langefeld, T. H. Beaty, S. S. Rich, E. R. Bleecker, N. J. Cox, and A. Collaborative Study on the Genetics of, "Genomewide screen and identification of gene-gene interactions for asthma-susceptibility loci in three U.S. populations: collaborative study on the genetics of asthma," *Am J Hum Genet,* vol. 68, pp. 1437-46, Jun 2001.

[177] S. H. Atkinson, S. M. Uyoga, E. Nyatichi, A. W. Macharia, G. Nyutu, C. Ndila, D. P. Kwiatkowski, K. A. Rockett, and T. N. Williams, "Epistasis between the haptoglobin common variant and alpha+thalassemia influences risk of severe malaria in Kenyan children," *Blood,* vol. 123, pp. 2008-16, Mar 27 2014.

[178] D. P. Kwiatkowski, "How malaria has affected the human genome and what human genetics can teach us about malaria," *Am J Hum Genet,* vol. 77, pp. 171-92, Aug 2005.

[179] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge *et. al,* "The genetic landscape of a cell," *Science,* vol. 327, pp. 425-31, Jan 22 2010.

[180] S. M. Taylor, C. M. Parobek, and R. M. Fairhurst, "Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis," *Lancet Infect Dis,* vol. 12, pp. 457-68, Jun 2012.

[181] T. E. Wellems and R. M. Fairhurst, "An evolving picture of the interactions between malaria parasites and their host erythrocytes," *Cell Res,* vol. 22, pp. 453-6, Mar 2012.

[182] C. M. Mbogo, J. M. Mwangangi, J. Nzovu, W. Gu, G. Yan, J. T. Gunter, C. Swalm, J. Keating, J. L. Regens, J. I. Shililu, J. I. Githure, and J. C. Beier, "Spatial and temporal heterogeneity of Anopheles mosquitoes and Plasmodium falciparum transmission along the Kenyan coast," *Am J Trop Med Hyg,* vol. 68, pp. 734-42, Jun 2003.

[183] C. N. Mbogo, R. W. Snow, C. P. Khamala, E. W. Kabiru, J. H. Ouma, J. I. Githure, K. Marsh, and J. C. Beier, "Relationships between Plasmodium falciparum transmission by vector populations and the incidence of severe disease at nine sites on the Kenyan coast," *Am J Trop Med Hyg,* vol. 52, pp. 201-6, Mar 1995.

[184] C. N. Mbogo, R. W. Snow, E. W. Kabiru, J. H. Ouma, J. I. Githure, K. Marsh, and J. C. Beier, "Low-level *Plasmodium falciparum* transmission and the incidence of severe malaria infections on the Kenyan coast," *Am J Trop Med Hyg,* vol. 49, pp. 245-53, Aug 1993.

[185] W. P. O'Meara, P. Bejon, T. W. Mwangi, E. A. Okiro, N. Peshu, R. W. Snow, C. R. Newton, and K. Marsh, "Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya," *Lancet,* vol. 372, pp. 1555-62, Nov 1 2008.

[186] W. P. O'Meara, T. W. Mwangi, T. N. Williams, F. E. McKenzie, R. W. Snow, and K. Marsh, "Relationship between exposure, clinical malaria, and age in an area of changing transmission intensity," *Am J Trop Med Hyg,* vol. 79, pp. 185-91, Aug 2008.

[187] J. A. Scott, E. Bauni, J. C. Moisi, J. Ojal, H. Gatakaa, C. Nyundo, C. S. Molyneux, F. Kombe, B. Tsofa, K. Marsh, N. Peshu, and T. N. Williams, "Profile: The Kilifi Health and Demographic Surveillance System (KHDSS)," *Int J Epidemiol,* vol. 41, pp. 650-7, Jun 2012.

[188] INDEPTH Network. *An international Network of field sites with continuous Demographic Evaluation of Populations and their Health in developing countries.* Available: http://www.indepth-network.org

[189] C. G. Nevill, E. S. Some, V. O. Mung'ala, W. Mutemi, L. New, K. Marsh, C. Lengeler, and R. W. Snow, "Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast," *Trop Med Int Health,* vol. 1, pp. 139-46, Apr 1996.

[190] N. A. Eid, A. A. Hussein, A. M. Elzein, H. S. Mohamed, K. A. Rockett, D. P.Kwiatkowski, and M. E. Ibrahim, "Candidate malaria susceptibility /protective SNPs in hospital and population-based studies: the effect of sub-structuring," *Malar J,* vol. 9, p. 119, 2010.

[191] R. W. Snow, J. R. Armstrong, D. Forster, M. T. Winstanley, V. M. Marsh, C. R. Newton, C. Waruiru, I. Mwangi, P. A. Winstanley, and K. Marsh, "Childhood deaths in Africa: uses and limitations of verbal autopsies," *Lancet,* vol. 340, pp. 351-5, Aug 8 1992.

[192] E. A. Okiro, S. I. Hay, P. W. Gikandi, S. K. Sharif, A. M. Noor, N. Peshu, K. Marsh, and R. W. Snow, "The decline in paediatric malaria admissions on the coast of Kenya," *Malar J,* vol. 6, p. 151, 2007.

[193] E. A. Okiro, V. A. Alegana, A. M. Noor, J. J. Mutheu, E. Juma, and R. W. Snow, "Malaria paediatric hospitalization between 1999 and 2008 across Kenya," *BMC Med,* vol. 7, p. 75, 2009.

[194] C. Ndila, E. Bauni, V. Nyirongo, G. Mochamah, A. Makazi, P. Kosgei, G. Nyutu, A. Macharia, S. Kapesa, P. Byass, and T. N. Williams, "Verbal autopsy as a tool for identifying children dying of sickle cell disease: a validation study conducted in Kilifi district, Kenya," *BMC Med,* vol. 12, p. 65, 2014.

[195] K. Maitland, A. Pamba, C. R. Newton, and M. Levin, "Response to volume resuscitation in children with severe malaria," *Pediatr Crit Care Med,* vol. 4, pp. 426-31, Oct 2003.

[196] T. N. Williams, S. Uyoga, A. Macharia, C. Ndila, C. F. McAuley, D. H. Opi, S. Mwarumba, J. Makani, A. Komba, M. N. Ndiritu, S. K. Sharif, K. Marsh, J. A. Berkley, and J. A. Scott, "Bacteraemia in Kenyan children with sickle-cell anaemia: a retrospective cohort and case-control study," *Lancet,* vol. 374, pp. 1364-70, Oct 17 2009.

[197] L. Zhang, X. Cui, K. Schmitt, R. Hubert, W. Navidi, and N. Arnheim, "Whole genome amplification from a single cell: implications for genetic analysis," *Proc Natl Acad Sci U S A,* vol. 89, pp. 5847-51, Jul 1 1992.

[198] J. M. Gonzalez, M. C. Portillo, and C. Saiz-Jimenez, "Multiple displacement amplification as a pre-polymerase chain reaction (pre-PCR) to process difficult to amplify samples and low copy number sequences from natural environments," *Environ Microbiol,* vol. 7, pp. 1024-8, Jul 2005.

[199] P. Ross, L. Hall, I. Smirnov, and L. Haff, "High level multiplex genotyping by MALDI-TOF mass spectrometry," *Nat Biotechnol,* vol. 16, pp. 1347-51, Dec 1998.

[200] "A global network for investigating the genomic epidemiology of malaria," *Nature,* vol. 456, pp. 732-7, Dec 11 2008.

[201] O. Delaneau, J. F. Zagury, and J. Marchini, "Improved whole-chromosome phasing for disease and population genetic studies," *Nat Methods,* vol. 10, pp. 5-6, Jan 2013.

[202] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet,* vol. 5, p. e1000529, Jun 2009.

[203] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics,* vol. 29, pp. 311-22, Sep 20 1995.

[204] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nat Rev Genet,* vol. 10, pp. 392-404, Jun 2009.

[205] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet,* vol. 81, pp. 559-75, Sep 2007.

[206] M. Dorigo and L. M. Gambardella, "Ant colonies for the travelling salesman problem," *Biosystems,* vol. 43, pp. 73-81, 1997.

[207] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Res Notes,* vol. 3, p. 117, 2010.

[208] J. Shang, J. Zhang, Y. Sun, D. Liu, D. Ye, and Y. Yin, "Performance analysis of novel methods for detecting epistasis," *BMC Bioinformatics,* vol. 12, p. 475, 2011.

[209] T. K. Rice, N. J. Schork, and D. C. Rao, "Methods for handling multiple testing," *Adv Genet,* vol. 60, pp. 293-308, 2008.

[210] A. Manjurano, T. G. Clark, B. Nadjm, G. Mtove, H. Wangai, N. Sepulveda, S. G. Campino, C. Maxwell, R. Olomi, K. R. Rockett, A. Jeffreys, C. MalariaGen, E. M. Riley, H. Reyburn, and C. Drakeley, "Candidate human genetic polymorphisms and severe malaria in a Tanzanian population," *PLoS One,* vol. 7, p. e47463, 2012.

[211] O. Toure, S. Konate, S. Sissoko, A. Niangaly, A. Barry, A. H. Sall, E. Diarra, B. Poudiougou, N. Sepulveda, S. Campino, K. A. Rockett, T. G. Clark, M. A. Thera, O. Doumbo, and G. E. N. C. Collaboration with The Malaria, "Candidate polymorphisms and severe malaria in a Malian population," *PLoS One,* vol. 7, p. e43987, 2012.

[212] S. da Silva Santos, T. G. Clark, S. Campino, M. C. Suarez-Mutis, K. A. Rockett, D. P. Kwiatkowski, and O. Fernandes, "Investigation of host candidate malaria-associated risk/protective SNPs in a Brazilian Amazonian population," *PLoS One,* vol. 7, p. e36692, 2012.

[213] N. Malaria Genomic Epidemiology and N. Malaria Genomic Epidemiology, "Genetic resistance to severe malaria is associated with a locus of ancient balancing selection," *under review,* 2015.

[214] F. J. Fowkes, S. J. Allen, A. Allen, M. P. Alpers, D. J. Weatherall, and K. P. Day, "Increased microerythrocyte count in homozygous alpha(+)-thalassaemia contributes to protection against severe malarial anaemia," *PLoS Med,* vol. 5, p. e56, Mar 18 2008.

[215] A. Dzeing-Ella, P. C. Nze Obiang, R. Tchoua, T. Planche, B. Mboza, M. Mbounja, U. Muller-Roemer, J. Jarvis, E. Kendjo, E. Ngou-Milama, P. G. Kremsner, S. Krishna, and M. Kombila, "Severe falciparum malaria in Gabonese children: clinical and laboratory features," *Malar J,* vol. 4, p. 1, Jan 9 2005.

[216] I. Sanou, J. Pare, S. Traore, D. Modiano, K. L. Kam, J. Kabore, L. Lamizana, S. A. Sawadogo, and T. R. Guiguemde, "[Clinical signs of severe malaria in a pediatric hospital in Ouagadougou]," *Sante,* vol. 7, pp. 13-7, Jan-Feb 1997.

[217] B. M. Greenwood, A. K. Bradley, A. M. Greenwood, P. Byass, K. Jammeh, K. Marsh, S. Tulloch, F. S. Oldfield, and R. Hayes, "Mortality and morbidity from malaria among children in a rural area of The Gambia, West Africa," *Trans R Soc Trop Med Hyg,* vol. 81, pp. 478-86, 1987.

[218] D. Schellenberg, C. Menendez, E. Kahigwa, F. Font, C. Galindo, C. Acosta, J. A. Schellenberg, J. J. Aponte, J. Kimario, H. Urassa, H. Mshinda, M. Tanner, and P. Alonso, "African children with malaria in an area of intense Plasmodium falciparum transmission: features on admission to the hospital and risk factors for death," *Am J Trop Med Hyg,* vol. 61, pp. 431-8, Sep 1999.

[219] S. J. Allen, A. O'Donnell, N. D. Alexander, and J. B. Clegg, "Severe malaria in children in Papua New Guinea," *QJM,* vol. 89, pp. 779-88, Oct 1996.

[220] D. Waller, S. Krishna, J. Crawley, K. Miller, F. Nosten, D. Chapman, F. O. ter Kuile, C. Craddock, C. Berry, P. A. Holloway, and et al., "Clinical features and outcome of severe malaria in Gambian children," *Clin Infect Dis,* vol. 21, pp. 577-87, Sep 1995.

[221] S. Krishna, D. W. Waller, F. ter Kuile, D. Kwiatkowski, J. Crawley, C. F. Craddock, F. Nosten, D. Chapman, D. Brewster, P. A. Holloway, and et al., "Lactic acidosis and hypoglycaemia in children with severe malaria: pathophysiological and prognostic significance," *Trans R Soc Trop Med Hyg,* vol. 88, pp. 67-73, Jan-Feb 1994.

[222] H. Ackerman, S. Usen, M. Jallow, F. Sisay-Joof, M. Pinder, and D. P. Kwiatkowski, "A comparison of case-control and family-based association methods: the example of sickle-cell and malaria," *Ann Hum Genet,* vol. 69, pp. 559-65, Sep 2005.

[223] T. N. Williams, T. W. Mwangi, D. J. Roberts, N. D. Alexander, D. J. Weatherall, S. Wambua, M. Kortok, R. W. Snow, and K. Marsh, "An immune basis for malaria protection by the sickle cell trait," *PLoS Med,* vol. 2, p. e128, May 2005.

[224] A. E. Fry, M. J. Griffiths, S. Auburn, M. Diakite, J. T. Forton, A. Green, A. Richardson, J. Wilson, M. Jallow, F. Sisay-Joof, M. Pinder *et.al,* "Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria," *Hum Mol Genet,* vol. 17, pp. 567-76, Feb 15 2008.

[225] A. C. Allison and D. F. Clyde, "Malaria in African children with deficient erythrocyte glucose-6-phosphate dehydrogenase," *Br Med J,* vol. 1, pp. 1346-9, May 13 1961.

[226] H. M. Gilles, K. A. Fletcher, R. G. Hendrickse, R. Lindner, S. Reddy, and N. Allan, "Glucose-6-phosphate-dehydrogenase deficiency, sickling, and malaria in African children in South Western Nigeria," *Lancet,* vol. 1, pp. 138-40, Jan 21 1967.

[227] T. Butler, "G-6-PD deficiency and malaria in Black Americans in Vietnam," *Mil Med,* vol. 138, pp. 153-5, Mar 1973.

[228] S. K. Martin, L. H. Miller, D. Alling, V. C. Okoye, G. J. Esan, B. O. Osunkoya, and M. Deane, "Severe malaria and glucose-6-phosphate-dehydrogenase deficiency: a reappraisal of the malaria/G-6-P.D. hypothesis," *Lancet,* vol. 1, pp. 524-6, Mar 10 1979.

[229] S. S. Shah, A. Macharia, J. Makale, S. Uyoga, K. Kivinen, R. Craik, C. Hubbart, T. E. Wellems, K. A. Rockett, D. P. Kwiatkowski, and T. N. Williams, "Genetic determinants of glucose-6-phosphate dehydrogenase activity in Kenya," *BMC Med Genet,* vol. 15, p. 93, 2014.

[230] C. Timmann, T. Thye, M. Vens, J. Evans, J. May, C. Ehmen, J. Sievertsen, B. Muntau, G. Ruge, W. Loag, D. Ansong, S. Antwi, E. Asafo-Adjei *et.al,* "Genome-wide association study indicates two novel resistance loci for severe malaria," *Nature,* vol. 489, pp. 443-6, Sep 20 2012.

[231] T. Planche, A. Dzeing, E. Ngou-Milama, M. Kombila, and P. W. Stacpoole, "Metabolic complications of severe malaria," *Curr Top Microbiol Immunol,* vol. 295, pp. 105-36, 2005.

[232] C. V. Plowe, J. F. Cortese, A. Djimde, O. C. Nwanyanwu, W. M. Watkins, P. A. Winstanley, J. G. Estrada-Franco, R. E. Mollinedo, J. C. Avila, J. L. Cespedes, D. Carter, and O. K. Doumbo, "Mutations in Plasmodium falciparum dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance," *J Infect Dis,* vol. 176, pp. 1590-6, Dec 1997.

[233] L. H. Miller, D. I. Baruch, K. Marsh, and O. K. Doumbo, "The pathogenic basis of malaria," *Nature,* vol. 415, pp. 673-9, Feb 7 2002.

[234] T. E. Taylor, A. Borgstein, and M. E. Molyneux, "Acid-base status in paediatric Plasmodium falciparum malaria," *Q J Med,* vol. 86, pp. 99-109, Feb 1993.

[235] F. P. Mockenhaupt, S. Ehrhardt, J. Burkhardt, S. Y. Bosomtwe, S. Laryea, S. D. Anemana, R. N. Otchwemah, J. P. Cramer, E. Dietz, S. Gellert, and U. Bienzle, "Manifestation and outcome of severe malaria in children in northern Ghana," *Am J Trop Med Hyg,* vol. 71, pp. 167-72, Aug 2004.

[236] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nat Genet,* vol. 37, pp. 413-7, Apr 2005.

[237] D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon, "Two-stage two-locus models in genome-wide association," *PLoS Genet,* vol. 2, p. e157, Sep 22 2006.

[238] C. Yang, X. Wan, Z. He, Q. Yang, H. Xue, and W. Yu, "The choice of null distributions for detecting gene-gene interactions in genome-wide association studies," *BMC Bioinformatics,* vol. 12 Suppl 1, p. S26, 2011.

[239] R. C. Culverhouse, "A comparison of methods sensitive to interactions with small main effects," *Genet Epidemiol,* vol. 36, pp. 303-11, May 2012.

[240] A. M. Molinaro, N. Carriero, R. Bjornson, P. Hartge, N. Rothman, and N. Chatterjee, "Power of data mining methods to detect genetic associations and interactions," *Hum Hered,* vol. 72, pp. 85-97, 2011.

[241] C. C. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan, "Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 8, pp. 1580-91, Nov-Dec 2011.

[242] M. Garcia-Magarinos, I. Lopez-de-Ullibarri, R. Cao, and A. Salas, "Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction," *Ann Hum Genet,* vol. 73, pp. 360-9, May 2009.

[243] K. Kapur, T. Schupbach, I. Xenarios, Z. Kutalik, and S. Bergmann, "Comparison of strategies to detect epistasis from eQTL data," *PLoS One,* vol. 6, p. e28415, 2011.

[244] S. Winham, C. Wang, and A. A. Motsinger-Reif, "A comparison of multifactor dimensionality reduction and L1-penalized regression to identify gene-gene interactions in genetic association studies," *Stat Appl Genet Mol Biol,* vol. 10, p. Article 4, 2011.

[245] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Res,* vol. 33, pp. D433-7, Jan 1 2005.

[246] N. B. Larson and D. J. Schaid, "A kernel regression approach to gene-gene interaction detection for case-control studies," *Genet Epidemiol,* vol. 37, pp. 695-703, Nov 2013.

[247] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum Hered,* vol. 56, pp. 73-82, 2003.

[248] M. Chen, J. Cho, and H. Zhao, "Detecting epistatic SNPs associated with complex diseases via a Bayesian classification tree search method," *Ann Hum Genet,* vol. 75, pp. 112-21, Jan 2011.

[249] W. H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," *Nat Rev Genet,* vol. 15, pp. 722-33, Nov 2014.

[250] F. M. Brennan and M. Feldmann, "Cytokines in autoimmunity," *Curr Opin Immunol,* vol. 8, pp. 872-7, Dec 1996.

[251] M. Steffens, T. Becker, T. Sander, R. Fimmers, C. Herold, D. A. Holler, C. Leu, S. Herms, S. Cichon, B. Bohn, T. Gerstner, M. Griebel, M. M. Nothen, T. F. Wienker, and M. P. Baur, "Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests," *Hum Hered,* vol. 69, pp. 268-84, 2010.

[252] M. Emily, T. Mailund, J. Hein, L. Schauser, and M. H. Schierup, "Using biological networks to search for interacting loci in genome-wide association studies," *Eur J Hum Genet,* vol. 17, pp. 1231-40, Oct 2009.

[253] E. J. Rogers, A. C. Shone, S. Alonso, C. A. May, and J. A. Armour, "Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes," *Hum Mol Genet,* vol. 9, pp. 2675-81, Nov 1 2000.

[254] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *Am J Hum Genet,* vol. 68, pp. 978-89, Apr 2001.

[255] L. Kruglyak, "Prospects for whole-genome linkage disequilibrium mapping of common disease genes," *Nat Genet,* vol. 22, pp. 139-44, Jun 1999.

[256] S. E. Antonarakis, C. D. Boehm, P. J. Giardina, and H. H. Kazazian, Jr., "Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster," *Proc Natl Acad Sci U S A,* vol. 79, pp. 137-41, Jan 1982.

[257] D. Labie, R. Srinivas, O. Dunda, C. Dode, C. Lapoumeroulie, V. Devi, S. Devi, K. Ramasami, J. Elion, R. Ducrocq, and et al., "Haplotypes in tribal Indians bearing the sickle gene: evidence for the unicentric origin of the beta S mutation and the unicentric origin of the tribal populations of India," *Hum Biol,* vol. 61, pp. 479-91, Aug 1989.

[258] C. Lapoumeroulie, O. Dunda, R. Ducrocq, G. Trabuchet, M. Mony-Lobe, J. M. Bodo, P. Carnevale, D. Labie, J. Elion, and R. Krishnamoorthy, "A novel sickle cell mutation of yet another origin in Africa: the Cameroon type," *Hum Genet,* vol. 89, pp. 333-7, May 1992.

[259] E. T. Wood, D. A. Stover, M. Slatkin, M. W. Nachman, and M. F. Hammer, "The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria," *Am J Hum Genet,* vol. 77, pp. 637-42, Oct 2005.

[260] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," *Malawi Med J,* vol. 24, pp. 69-71, Sep 2012.

[261] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, *et.al* "Detecting recent positive selection in the human genome from haplotype structure," *Nature,* vol. 419, pp. 832-7, Oct 24 2002.

[262] M. A. Saunders, M. F. Hammer, and M. W. Nachman, "Nucleotide variability at G6pd and the signature of malarial selection in humans," *Genetics,* vol. 162, pp. 1849-61, Dec 2002.

[263] J. Ohashi, I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham, S. Looareesuwan, A. G. Clark, and K. Tokunaga, "Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection," *Am J Hum Genet,* vol. 74, pp. 1198-208, Jun 2004.

[264] E. M. Leffler, Z. Gao, S. Pfeifer, L. Segurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J. D. Wall, G. Sella, P. Donnelly, G. McVean, and M. Przeworski, "Multiple instances of ancient balancing selection shared between humans and chimpanzees," *Science,* vol. 339, pp. 1578-82, Mar 29 2013.

[265] A. C. Allison, "Notes on sickle-cell polymorphism," *Ann Hum Genet,* vol. 19, pp. 39-51, Jul 1954.

[266] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, *et.al,* "Genetic signatures of strong recent positive selection at the lactase gene," *Am J Hum Genet,* vol. 74, pp. 1111-20, Jun 2004.

[267] S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, *et. al* "Convergent adaptation of human lactase persistence in Africa and Europe," *Nat Genet,* vol. 39, pp. 31-40, Jan 2007.

[268] J. Pagnier, J. G. Mears, O. Dunda-Belkhodja, K. E. Schaefer-Rego, C. Beldjord, R. L. Nagel, and D. Labie, "Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa," *Proc Natl Acad Sci U S A,* vol. 81, pp. 1771-3, Mar 1984.

[269] J. Flint, R. M. Harding, A. J. Boyce, and J. B. Clegg, "The population genetics of the haemoglobinopathies," *Baillieres Clin Haematol,* vol. 6, pp. 215-62, Mar 1993.

[270] J. Flint, R. M. Harding, A. J. Boyce, and J. B. Clegg, "The population genetics of the haemoglobinopathies," *Baillieres Clin Haematol,* vol. 11, pp. 1-51, Mar 1998.

[271] S. H. Orkin and H. H. Kazazian, Jr., "The mutation and polymorphism of the human beta-globin gene and its surrounding DNA," *Annu Rev Genet,* vol. 18, pp. 131-71, 1984.

[272] J. S. Wainscoat, E. Kanavakis, D. J. Weatherall, J. Walker, M. Holmes-Seidle, M. Bobrow, and A. B. Donnison, "Regional localisation of the human alpha-globin genes," *Lancet,* vol. 2, pp. 301-2, Aug 8 1981.

[273] S. M. Williams, J. A. Canter, D. C. Crawford, J. H. Moore, M. D. Ritchie, and J. L. Haines, "Problems with genome-wide association studies," *Science,* vol. 316, pp. 1840-2, Jun 29 2007.

[274] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, *et. al,* "Finding the missing heritability of complex diseases," *Nature,* vol. 461, pp. 747-53, Oct 8 2009.

[275] E. S. Lander and D. Botstein, "Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms," *Proc Natl Acad Sci U S A,* vol. 83, pp. 7353-7, Oct 1986.

[276] J. Dupuis, P. O. Brown, and D. Siegmund, "Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent," *Genetics,* vol. 140, pp. 843-56, Jun 1995.

[277] H. K. Tang and D. Siegmund, "Mapping multiple genes for quantitative or complex traits," *Genet Epidemiol,* vol. 22, pp. 313-27, Apr 2002.

[278] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics,* vol. 55, pp. 997-1004, Dec 1999.

[279] R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke *et.al*, "LocusZoom: regional visualization of genome-wide association scan results," *Bioinformatics,* vol. 26, pp. 2336-7, Sep 15 2010.

[280] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. I. de Bakker, "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap," *Bioinformatics,* vol. 24, pp. 2938-9, Dec 15 2008.

[281] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat Rev Genet,* vol. 9, pp. 356-69, May 2008.

[282] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nat Rev Genet,* vol. 7, pp. 781-91, Oct 2006.

[283] O. Carlborg and C. S. Haley, "Epistasis: too often neglected in complex trait studies?," *Nat Rev Genet,* vol. 5, pp. 618-25, Aug 2004.

[284] W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models," *Hum Hered,* vol. 50, pp. 334-49, Nov-Dec 2000.

[285] S. Purcell and P. C. Sham, "Epistasis in quantitative trait locus linkage analysis: interaction or main effect?," *Behav Genet,* vol. 34, pp. 143-52, Mar 2004.

[286] T. Iinuma, T. Aoki, K. Arasaki, H. Hirose, A. Yamamoto, R. Samata, H. P. Hauri, N. Arimitsu, M. Tagaya, and K. Tani, "Role of syntaxin 18 in the organization of endoplasmic reticulum subdomains," *J Cell Sci,* vol. 122, pp. 1680-90, May 15 2009.

[287] K. Hatsuzawa, H. Hirose, K. Tani, A. Yamamoto, R. H. Scheller, and M. Tagaya, "Syntaxin 18, a SNAP receptor that functions in the endoplasmic reticulum, intermediate compartment, and cis-Golgi vesicle trafficking," *J Biol Chem,* vol. 275, pp. 13713-20, May 5 2000.

[288] T. Bassett, B. Harpur, H. Y. Poon, K. H. Kuo, and C. H. Lee, "Effective stimulation of growth in MCF-7 human breast cancer cells by inhibition of syntaxin18 by external guide sequence and ribonuclease P," *Cancer Lett,* vol. 272, pp. 167-75, Dec 8 2008.

[289] H. J. Cordell, J. Bentham, A. Topf, D. Zelenika, S. Heath, C. Mamasoula, C. Cosgrove, G. Blue, *et. al*, "Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16," *Nat Genet,* vol. 45, pp. 822-4, Jul 2013.

[290] A. C. Moss, G. Lawlor, D. Murray, D. Tighe, S. F. Madden, A. M. Mulligan, C. O. Keane, H. R. Brady, P. P. Doran, and P. MacMathuna, "ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion," *Biochem Biophys Res Commun,* vol. 345, pp. 216-21, Jun 23 2006.

[291] J. Leyden, D. Murray, A. Moss, M. Arumuguma, E. Doyle, G. McEntee, C. O'Keane, P. Doran, and P. MacMathuna, "Net1 and Myeov: computationally identified mediators of gastric cancer," *Br J Cancer,* vol. 94, pp. 1204-12, Apr 24 2006.

[292] A. Gyenesei, J. Moody, C. A. Semple, C. S. Haley, and W. H. Wei, "High-throughput analysis of epistasis in genome-wide association studies with BiForce," *Bioinformatics,* vol. 28, pp. 1957-64, Aug 1 2012.

[293] M. Ueki and H. J. Cordell, "Improved statistics for genome-wide interaction analysis," *PLoS Genet,* vol. 8, p. e1002625, 2012.

[294] L. Hsu, S. Jiao, J. Y. Dai, C. Hutter, U. Peters, and C. Kooperberg, "Powerful cocktail methods for detecting genome-wide gene-environment interaction," *Genet Epidemiol,* vol. 36, pp. 183-94, Apr 2012.

[295] K. V. Steen, "Travelling the world of gene-gene interactions," *Brief Bioinformatics,* vol. 13, pp. 1-19, Jan 2012.

[296] A. V. Hill, "Evolution, revolution and heresy in the genetics of infectious disease susceptibility," *Philos Trans R Soc Lond B Biol Sci,* vol. 367, pp. 840-9, Mar 19 2012.

[297] A. Manjurano, N. Sepulveda, B. Nadjm, G. Mtove, H. Wangai, C. Maxwell, R. Olomi, H. Reyburn, C. J. Drakeley, E. M. Riley, T. G. Clark, and G. E. N. in Collaboration With Malaria, "USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania," *J Infect Dis,* Mar 24 2015.

[298] J. Ohashi, I. Naka, J. Patarapotikul, H. Hananantachai, S. Looareesuwan, and K. Tokunaga, "Absence of association between the allele coding methionine at position 29 in the N-terminal domain of ICAM-1 (ICAM-1(Kilifi)) and severe malaria in the northwest of Thailand," *Jpn J Infect Dis,* vol. 54, pp. 114-6, Jun 2001.

[299] U. Bienzle, T. A. Eggelte, L. A. Adjei, E. Dietz, S. Ehrhardt, J. P. Cramer, R. N. Otchwemah, and F. P. Mockenhaupt, "Limited influence of haptoglobin genotypes on severe malaria in Ghanaian children," *Trop Med Int Health,* vol. 10, pp. 668-71, Jul 2005.

[300] S. R. Collins, K. M. Miller, N. L. Maas, A. Roguev, J. Fillingham, C. S. Chu, M. Schuldiner, M. Gebbia, J. Recht, M. Shales, H. Ding, H. Xu, *et. al,* "Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map," *Nature,* vol. 446, pp. 806-10, Apr 12 2007.